REAL ANALYSIS AND FOUNDATIONS

FOURTH EDITION

Steven G. Krantz

REAL ANALYSIS AND FOUNDATIONS
FOURTH EDITION

TEXTBOOKS in MATHEMATICS

Series Editors: Al Boggess and Ken Rosen

PUBLISHED TITLES

ABSTRACT ALGEBRA: A GENTLE INTRODUCTION
Gary L. Mullen and James A. Sellers

ABSTRACT ALGEBRA: AN INTERACTIVE APPROACH, SECOND EDITION
William Paulsen

ABSTRACT ALGEBRA: AN INQUIRY-BASED APPROACH
Jonathan K. Hodge, Steven Schlicker, and Ted Sundstrom

ADVANCED LINEAR ALGEBRA
Hugo Woerdeman

ADVANCED LINEAR ALGEBRA
Nicholas Loehr

ADVANCED LINEAR ALGEBRA, SECOND EDITION
Bruce Cooperstein

APPLIED ABSTRACT ALGEBRA WITH MAPLE™ AND MATLAB®, THIRD EDITION
Richard Klima, Neil Sigmon, and Ernest Stitzinger

APPLIED DIFFERENTIAL EQUATIONS: THE PRIMARY COURSE
Vladimir Dobrushkin

A BRIDGE TO HIGHER MATHEMATICS
Valentin Deaconu and Donald C. Pfaff

COMPUTATIONAL MATHEMATICS: MODELS, METHODS, AND ANALYSIS WITH MATLAB® AND MPI, SECOND EDITION
Robert E. White

A COURSE IN DIFFERENTIAL EQUATIONS WITH BOUNDARY VALUE PROBLEMS, SECOND EDITION
Stephen A. Wirkus, Randall J. Swift, and Ryan Szypowski

A COURSE IN ORDINARY DIFFERENTIAL EQUATIONS, SECOND EDITION
Stephen A. Wirkus and Randall J. Swift

DIFFERENTIAL EQUATIONS: THEORY, TECHNIQUE, AND PRACTICE, SECOND EDITION
Steven G. Krantz

PUBLISHED TITLES CONTINUED

REAL ANALYSIS AND FOUNDATIONS
FOURTH EDITION

Steven G. Krantz

Washington University
St. Louis, Missouri, USA

**Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com**

**and the CRC Press Web site at
http://www.crcpress.com**

To Stan Philipp, who taught me real analysis.
And to Walter Rudin, who wrote the books from which I learned.

Table of Contents

Preface to the Fourth Edition

The first three editions of this book have become well established, and have attracted a loyal readership. Students find the book to be concise, accessible, and complete. Instructors find the book to be clear, authoritative, and dependable. We continue to listen to our readership and to revise the book to make it more meaningful for new generations of mathematics students.

This new edition retains many of the basic features of the earlier versions. We still cover sequences, series, functions, limits of sequences and series of functions, differentiation theory, and integration theory. The theory of functions of several variables is explored. And introductions to Fourier analysis and differential equations is included to make the book timely and relevant.

In this new edition we endeavor to make the book accessible to a broader audience. We do not want this to be perceived as a "high level" text. Therefore we include more explanation, more elementary examples, and we stepladder the exercises. We update and clarify the figures. We make the sections more concise, and omit technical details which are not needed for a solid and basic understanding of the key ideas.

The book assumes that the student has a solid background in calculus, and at least some experience with mathematical reasoning. We do not assume that the student knows topology or linear algebra. A typical student in a course using this book would be a junior with a major in an analytical science.

In the same spirit, we have eliminated Chapter 13 on advanced topics and Chapter 14 on normed linear spaces. These are very attractive sets of ideas, but are probably best treated in a more advanced course.

We have updated and augmented the multivariable material in order to bring out the geometric nature of the topic. The figures are thus enhanced and fleshed out. Clearly functions of several variables is a suitable climax for this textbook.

We retain the chapter on Fourier analysis because it is timely and indicates applications of the key ideas in the book. Certainly wavelets are one of the

hot ideas in modern analysis, and this chapter gives us an opportunity to give the reader a taste of that subject. The chapter is self contained, and some instructors may choose to skip it without any loss of continuity.

The chapter on differential equations gives the student some hands-on experience with how analysis is applied to real-world questions. The differential equations that we treat are classical, but the ideas are still current.

Our goal with this new edition is to make the important but rather advanced subject of real analysis relevant and accessible to a broad audience of students with diverse backgrounds. Real analysis is not just for future mathematicians. It is a basic tool for all mathematical scientists, ranging from physicists to engineers to researchers in the medical profession. We want this text to be the generational touchstone for the subject and the go-to text for developing young scientists.

It is a pleasure to thank the several reviewers who contributed decisively to the clarity and correctness of my manuscript. As always, I thank my editor Robert Ross for his encouragement and support.

We anticipate useful feedback from the readership, and we look forward to the development of a new generation of young analysts.

<div align="right">

Steven G. Krantz
St. Louis, Missouri

</div>

Preface to the Third Edition

The enthusiastic reception that the first two editions of this book have received has been pleasing. We write this third edition with a view to making the text more flexible and useful for our readers.

In particular, we have endeavored to make the book more streamlined. Worth mention are these changes:

- The beginning chapter on set theory and logic is now an appendix. Many students will be familiar with this material, and can refer to this for review when needed. In this way we can get the students quickly into the guts of real analysis.

- The chapter on number systems has been shortened so that it now concentrates on the real numbers and the complex numbers. It is safe to assume that students are familiar with the natural numbers, the integers, and the rational numbers. The more elementary number systems are treated in Appendix I at the end of the book.

- Differential equations have always been the wellspring of analysis, and we have retained most of our applications to ordinary differential equations and partial differential equations. But we have removed the material on the method of characteristics, as it is ancillary.

- We have removed the chapter on wavelet theory, as it is truly beyond the scope of a typical real variables course.

- We have removed the material on measure theory because it is just too difficult.

- We have removed the material on differential forms as it is really best suited to a more geometric course.

There are still some nice chapters that, after the student has been through the basic material, offer some "dessert." These include a treatment of Riemann–Stieltjes integrals, a chapter on Fourier analysis, a chapter on metric spaces and applications, and a chapter on differential equations. We have added a chapter on normed linear spaces. After all, a look at infinite dimensional analysis is definitely a glimpse of future work.

Of course we have taken this opportunity to augment most every chapter with additional examples and exercises. Difficult or challenging exercises are still marked with a *. And we have corrected some errors and clarified some passages.

Instead of having exercise sets at the end of each chapter, we now have exercise sets at the end of each section. This will make the book more useful, and also help the students to key the exercises to appropriate text passages.

The student who works with this book will come away with a solid foundation in mathematical analysis and its applications. He/she will be ready for further exploration of measure theory, functional analysis, harmonic analysis, and beyond.

It is a particular pleasure to thank Zhongshan "Jason" Li and his colleagues at Georgia State University for their many edits and corrections. All this terrific information has definitely resulted in a better book.

Bob Stern has been my editor for many years, and has always been a trusted collaborator and friend. I thank him for our many books.

As always, we welcome responses and input from our readers.

— Steven G. Krantz
St. Louis, Missouri

Preface to the Second Edition

The book *Real Analysis and Foundations*, first published in 1991, is unique in several ways. It was the first book to attempt a bridge between the rather hard-edged classical books in the subject—like Walter Rudin's *Principles of Mathematical Analysis*—and the softer and less rigorous books of today. This book combines authority, rigor, and readability in a manner that makes the subject accessible to students while still teaching them the strict discourse of mathematics.

Real Analysis and Foundations was a timely book, and it has been a successful book. It is used not only in mathematics departments but also in economics and physics and engineering and finance programs. The book's wide acceptance speaks for itself. Since the volume has been in print for thirteen years, it seems that a new edition is long overdue.

Like much of classical mathematics, real analysis is a subject that is immutable. It has not changed appreciably for 150 years, and it is not about to change. But there are new ideas that build on the old ones, and the presentation can evolve as well. In this new edition, we propose to build on the basic ideas of Fourier analysis (Chapter 12) and to develop some of the new ideas about wavelets (Chapter 15). We will indicate applications of wavelets to the theory of signal processing.

We can also augment the Fourier-analytic theory with applications to ordinary differential equations, and even to some partial differential equations. Elliptic boundary value problems on the disc, and their interpretation in terms of steady-state heat flow, are a natural crucible for the applications of real analysis.

As part of our treatment of differential equations we present the method of power series, the method of characteristics, and the Picard existence and uniqueness theorem. These are lovely pieces of mathematics, and they also allow us to show how fundamental ideas like uniform convergence and power series are applied.

We will amplify the development of real analysis of several variables. After all, the real world is three-dimensional and we must have the tools of multi-variable analysis in order to attack the concrete engineering problems that arise in higher dimensions. We will present the rudiments of the Lebesgue integration theory, primarily as an invitation to further study. We will also present the basics of differential forms and integration on surfaces. We will give a brief treatment of Stokes's theorem and its variants.

The exercise sets are rich and robust. Each chapter has an extensive and diverse collection of problems. Difficult or challenging exercises are marked with a *.

Of course we have re-thought and developed all the exercise sets and all the examples in the book. We have added more figures. We have corrected the few errors that have arisen over the years, tightened up the statements and proofs of the theorems, and provided end-of-section appendices to help the student with review topics.

In sum, the second edition of *Real Analysis and Foundations* will be a new book—even more lively and more vital than the popular first edition. I am happy to express my gratitude to my editor Robert Stern, who made this publishing experience a smooth and happy one. I look forward to hearing remarks and criticisms from my readers, in hopes of making future editions of this book more accurate and more useful.

— Steven G. Krantz
St. Louis, Missouri

Preface to the First Edition

Overview

The subject of real analysis, or "advanced calculus," has a central position in undergraduate mathematics education. Yet, because of changes in the preparedness of students, and because of their early exposure to calculus (and therefore lack of exposure to certain other topics) in high school, this position has eroded. Students unfamiliar with the value of rigorous, axiomatic mathematics are ill-prepared for a traditional course in mathematical analysis.

Thus there is a need for a book that simultaneously introduces students to rigor, to the *need* for rigor, and to the subject of mathematical analysis. The correct approach, in my view, is not to omit important classical topics like the Weierstrass Approximation theorem and the Ascoli-Arzela theorem, but rather to find the simplest and most direct path to each. While mathematics should be written "for the record" in a deductive fashion, proceeding from axioms to special cases, this is *not* how it is learned. Therefore (for example) I *do* treat metric spaces (a topic that has lately been abandoned by many of the current crop of analysis texts). I do so not at first but rather at the end of the book as a method for unifying what has gone before. And I do treat Riemann–Stieltjes integrals, but only after first doing Riemann integrals. I develop real analysis gradually, beginning with treating sentential logic, set theory, and constructing the integers.

The approach taken here results, in a technical sense, in some repetition of ideas. But, again, this is how one learns. Every generation of students comes to the university, and to mathematics, with its own viewpoint and background. Thus I have found that the classic texts from which we learned mathematical analysis are often no longer suitable, or appear to be inaccessible, to the present crop of students. It is my hope that my text will be a suitable source for modern students to learn mathematical analysis. Unlike other authors, I do not believe that the subject has changed; therefore I have not altered the fundamental content of the course. But the point of view of the audience has changed, and I have written my book accordingly.

The current crop of real analysis texts might lead one to believe that real analysis is simply a rehash of calculus. Nothing could be further from the truth. But many of the texts written thirty years ago are simply too dry and austere for today's audience. My purpose here is to teach today's students the mathematics that I grew to love in a language that speaks to them.

Prerequisites

A student with a standard preparation in lower division mathematics—calculus and differential equations—has adequate preparation for a course based on this text. Many colleges and universities now have a "transitions" course that helps students develop the necessary mathematical maturity for an upper division course such as real analysis. I have taken the extra precaution of providing a mini-transitions course in my Chapters 1 and 2. Here I treat logic, basic set theory, methods of proof, and constructions of the number systems. Along the way, students learn about mathematical induction, equivalence classes, completeness, and many other basic constructs. In the process of reading these chapters, written in a rigorous but inviting fashion, the student should gain both a taste and an appreciation for the use of rigor. While many instructors will want to spend some class time with these two chapters, others will make them assigned reading and begin the course proper with Chapter 3.

How to Build a Course from This Text

Chapters 3 through 7 present a first course in real analysis. I begin with the simplest ideas—sequences of numbers—and proceed to series, topology (on the real line only), limits and continuity of functions, and differentiation of functions. The order of topics is similar to that in traditional books like *Principles of Mathematical Analysis* by Walter Rudin, but the treatment is more gentle. There are many more examples, and much more explanation. I do not shortchange the really interesting topics like compactness and connectedness. The exercise sets provide plenty of drill, in addition to the more traditional "Prove this, Prove that." If it is possible to obtain a simpler presentation by giving up some generality, I always opt for simplicity.

Today many engineers and physicists are required to take a term of real analysis. Chapters 3 through 7 are designed for that purpose. For the more mathematically inclined, this first course serves as an introduction to the more advanced topics treated in the second part of the book.

In Chapter 8 I give a rather traditional treatment of the integral. First the Riemann integral is covered, then the Riemann–Stieltjes integral. I am careful to establish the latter integral as the natural setting for the integration by parts theorem. I establish explicitly that series are a special case of the Riemann–Stieltjes integral. Functions of bounded variation are treated briefly and their utility in integration theory is explained.

The usual material on sequences and series of functions in Chapter 9 (*in-*

cluding uniform convergence) is followed by a somewhat novel chapter on "Special Functions." Here I give a rigorous treatment of the elementary transcendental functions as well as an introduction to the gamma function and its application to Stirling's formula. The chapter concludes with an invitation to Fourier series.

I feel strongly, based in part on my own experience as a student, that analysis of several variables is a tough nut the first time around. In particular, college juniors and seniors are not (except perhaps at the very best schools) ready for differential forms. Therefore my treatment of functions of several variables in Chapter 11 is brief, it is only in $\mathbb{R}^3$, and it excludes any reference to differential forms. The main interests of this chapter, from the student's point of view, are (i) that derivatives are best understood using linear algebra and matrices and (ii) that the inverse function theorem and implicit function theorem are exciting new ideas. There are many fine texts that cover differential forms and related material and the instructor who wishes to treat that material in depth should supplement my text with one of those.

Chapter 12 [now Chapter 14] is dessert. For I have waited until now to introduce the language of metric spaces. But now comes the power, for I prove and apply both the Baire category theorem and the Ascoli-Arzela theorem. This is a suitable finish to a year-long course on the elegance and depth of rigorous reasoning.

I would teach my second course in real analysis by covering all of Chapters 8 through 12. Material in Chapters 10 and 12 is easily omitted if time is short.

Audience

This book is intended for college juniors and seniors and some beginning graduate students. It addresses the same niche as the classic books of Apostol, Royden, and Rudin. However, the book is written for today's audience in today's style. All the topics which excited my sense of wonder as a student—the Cantor set, the Weierstrass nowhere differentiable function, the Weierstrass approximation theorem, the Baire category theorem, the Ascoli-Arzela theorem—are covered. They can be skipped by those teaching a course for which these topics are deemed inappropriate. But they give the subject real texture.

Acknowledgments

It is a pleasure to thank Marco Peloso for reading the entire manuscript of this book and making a number of useful suggestions and corrections. Responsibility for any remaining errors of course resides entirely with me.

Peloso also wrote the solutions manual, which certainly augments the usefulness of the book.

Peter L. Duren, Peter Haskell, Kenneth D. Johnson, and Harold R. Parks served as reviewers of the manuscript that was submitted to CRC Press. Their

comments contributed decisively to the clarity and correctness of many passages. I am also grateful to William J. Floyd for a number of helpful remarks.

Russ Hall of CRC Press played an instrumental and propitious role in recruiting me to write for this publishing house. Wayne Yuhasz, Executive Editor of CRC Press, shepherded the project through every step of the production process. Lori Pickert of Archetype, Inc. typeset the book in TEX. All of these good people deserve my sincere thanks for the high quality of the finished book.

— Steven G. Krantz
St. Louis, Missouri

Chapter 1

Number Systems

1.1 The Real Numbers

This is a book about analysis in the real number system. Such a study must be founded on a careful consideration of *what the real numbers are* and *how they are constructed*. In the present section we give a careful treatment of the real number system. In the next we consider the complex numbers.

We know from calculus that, for many purposes, the rational numbers are inadequate. It is important to work in a number system which is closed with respect to the operations we shall perform. This includes limiting operations. While the rationals are closed under the usual arithmetic operations (addition, subtraction, multiplication, division), they are *not* closed under the mathematical operation of taking *limits*. For instance, the sequence of rational numbers $3, 3.1, 3.14, 3.141, \ldots$ consists of terms that seem to be getting closer and closer together, *seem* to tend to some limit, and yet there is no rational number which will serve as a limit (of course it turns out that the limit is π—an "irrational" number).

We will now deal with the real number system, a system which contains all limits of sequences of rational numbers (as well as all limits of sequences of real numbers!). In fact our plan will be as follows: in this section we shall discuss all the requisite properties of the reals. The actual construction of the reals is rather subtle, and we shall put that in an Appendix to Section 1.1.

Definition 1.1 Let A be an ordered set and X a subset of A. The set X is called *bounded above* if there is an element $b \in A$ such that $x \leq b$ for all $x \in X$. We call the element b an *upper bound* for the set X.

EXAMPLE 1.2 Let $A = \mathbb{Q}$ (the rational numbers) with the usual ordering. The set $X = \{x \in \mathbb{Q} : 2 < x < 4\}$ is bounded above. For example, 15 is an upper bound for X. So are the numbers 12 and 4. It is interesting to observe that no element of this particular X can actually be an upper bound for X. The

number 4 is a good candidate, but 4 is not an element of X. In fact if $b \in X$ then $(b+4)/2 \in X$ and $b < (b+4)/2$, so b could not be an upper bound for X. □

It turns out that the most convenient way to formulate the notion that the real numbers have "no holes" (i.e. that all sequences which seem to be converging actually have something to converge to) is in terms of upper bounds.

Definition 1.3 Let A be an ordered set and X a subset of A. An element $b \in A$ is called a *least upper bound* (or *supremum*) for X if b is an upper bound for X and $b \leq b'$ for every upper bound b' for X. We denote the supremum of X by sup X. The supremum is also sometimes called the *least upper bound* and denoted by lub X.

By its very definition, if a least upper bound exists then it is unique. Notice that we *could have* phrased the definition as "The point b is the least upper bound for X if, whenever $c < b$, then c cannot be an upper bound for X."

EXAMPLE 1.4 In the last example, we considered the set X of rational numbers strictly between 2 and 4. We observed there that 4 is the least upper bound for X. Note that this least upper bound is not an element of the set X.

The set $Y = \{y \in \mathbb{Z} : -9 \leq y \leq 7\}$ has least upper bound 7. In this case, the least upper bound *is* an element of the set Y. □

Notice that we may define a lower bound for a subset of an ordered set in a fashion similar to that for an upper bound:

Definition 1.5 A point $\ell \in A$ is a lower bound for $X \subseteq A$ if $\ell \leq x$ for all $x \in X$. A *greatest lower bound* (or *infimum*) for X is then defined to be a lower bound c such that $c \geq c'$ for every lower bound c' for X. We denote the infimum of X by inf X. The infimum is also sometimes called the *greatest lower bound* and denoted by glb X.

As with the least upper bound, we may note that the definition of greatest lower bound could be phrased in this way: "The point c is the greatest lower bound for X if, whenever $e > c$, then e cannot be a lower bound for X."

EXAMPLE 1.6 The set $X = \{x \in \mathbb{Q} : 2 < x < 4\}$ in the last two examples has lower bounds $-20, 0, 1, 2$, for instance. The greatest lower bound is 2, which is *not* an element of the set.

The set $Y = \{y \in \mathbb{Z} : -9 \leq y \leq 7\}$ in the last example has lower bounds— among others—given by $-53, -22, -10, -9$. The number -9 is the greatest lower bound. It *is* an element of Y. □

The purpose that the real numbers will serve for us is as follows: they will contain the rationals, they will still be an ordered field (a field is a set with arithmetic operations $+$ and $\cdot$ —see the Appendix at the end of Section 1.1), and *every subset which has an upper bound will have a least upper bound*. [See [KRA1] for a thorough treatment of the concept of ordered field.] We formulate this result as a theorem.

Theorem 1.7 *There exists an ordered field* $\mathbb{R}$ *which* **(i)** *contains* $\mathbb{Q}$ *and* **(ii)** *has the property that any nonempty subset of* $\mathbb{R}$ *which has an upper bound has a least upper bound (in the number system* $\mathbb{R}$*).*

The last property described in this theorem is called the Least Upper Bound Property of the real numbers. As mentioned previously, this theorem will be proved in the Appendix to Section 1.1. Of course the least upper bound property, in and of itself, is something of a technicality. But we shall see that a great many interesting and powerful properties of the real numbers can be derived from it.

Now we begin to realize why it is so important to *construct* the number systems that we will use. We are endowing $\mathbb{R}$ with a great many properties. Why do we have any right to suppose that there exists a set with all these properties? We must produce one! We do so in the Appendix to Section 1.1.

Let us begin to explore the richness of the real numbers. The next theorem states a property which is not shared by the rationals. It is fundamental in its importance.

Theorem 1.8 *Let* x *be a positive real number. Then there is a positive real number* y *such that* $y^2 = y \cdot y = x$.

Proof: We will use throughout this proof the fact that, if $0 < a < b$ then $a^2 < b^2$.

Let
$$S = \{s \in \mathbb{R} : s > 0 \text{ and } s^2 < x\}.$$

Then S is not empty since $x/2 \in S$ if $x < 2$ and $1 \in S$ otherwise. Also S is bounded above since $x + 1$ is an upper bound for S. By Theorem 1.6, the set S has a least upper bound. Call it y. Obviously, $0 < \min\{x/2, 1\} \leq y$ hence y is positive. We claim that $y^2 = x$. To see this, we eliminate the other two possibilities.

If $y^2 < x$ then set $\epsilon = (x - y^2)/[4(x + 1)]$. Then $\epsilon > 0$ and

$$
\begin{aligned}
(y + \epsilon)^2 &= y^2 + 2 \cdot y \cdot \epsilon + \epsilon^2 \\
&= y^2 + 2 \cdot y \cdot \frac{x - y^2}{4(x + 1)} + \frac{x - y^2}{4(x + 1)} \cdot \frac{x - y^2}{4(x + 1)} \\
&< y^2 + 2 \cdot y \cdot \frac{x - y^2}{4y} + \frac{x - y^2}{4(x + 1)} \cdot \frac{x - y^2}{4(x + 1)} \\
&< y^2 + \frac{x - y^2}{2} + \frac{x - y^2}{4} \cdot \frac{x}{4x} \\
&< y^2 + (x - y^2) \\
&= x.
\end{aligned}
$$

Thus $y + \epsilon \in S$, and y cannot be an upper bound for S. This contradiction tells us that $y^2 \not< x$.

Similarly, if it were the case that $y^2 > x$ then we set $\epsilon = (y^2 - x)/[4(x+1)]$. A calculation like the one we just did (see Exercise **5**) then shows that $(y-\epsilon)^2 \geq x$. Hence $y - \epsilon$ is also an upper bound for S, and y is therefore not the *least* upper bound. This contradiction shows that $y^2 \not> x$.

The only remaining possibility is that $y^2 = x$.   □

Remark 1.9 The theorem tells us in particular that $\sqrt{2}$, $\sqrt{5}$, $\sqrt{8}$, $\sqrt{11}$, etc. all exist in the real number system. And each of these numbers is irrational (see Theorem A1.23 where it is shown that $\sqrt{2}$ is irrational). In fact the only square roots of integers that are *not* irrational are the square roots of the perfect squares $1, 4, 9, 16, 25, \ldots$.   □

A similar proof shows that, if n is a positive integer and x a positive real number, then there is a positive real number y such that $y^n = x$. Exercise **14** asks you to provide the details.

We next use the Least Upper Bound Property of the Real Numbers to establish two important qualitative properties of the Real Numbers:

Theorem 1.10 *The set $\mathbb{R}$ of real numbers satisfies the Archimedean Property:*

> *Let a and b be positive real numbers. Then there is a natural number n such that $na > b$.*

The set $\mathbb{Q}$ of rational numbers satisfies the following Density Property:

> *Let $c < d$ be real numbers. Then there is a rational number q with $c < q < d$.*

Proof: Suppose the Archimedean Property to be false. Then $S = \{na : n \in \mathbb{N}\}$ has b as an upper bound. Therefore S has a finite supremum β. Since $a > 0$, it follows that $\beta - a < \beta$. So $\beta - a$ is not an upper bound for S, and there must be a natural number n' such that $n' \cdot a > \beta - a$. But then $(n' + 1)a > \beta$, and β cannot be the supremum for S. This contradiction proves the first assertion.

For the second property, let $\lambda = d - c > 0$. By the Archimedean Property, choose a positive integer N such that $N \cdot \lambda > 1$. Again the Archimedean Property gives a natural number P such that $P > N \cdot c$ and another Q such that $Q > -N \cdot c$. Thus we see that Nc falls between the integers $-Q$ and P; therefore there must be an integer M between $-Q$ and P such that

$$M - 1 \leq Nc < M.$$

Thus $c < M/N$. Also

$$M \leq Nc + 1 \quad \text{hence} \quad \frac{M}{N} \leq c + \frac{1}{N} < c + \lambda = d.$$

So M/N is a rational number lying between c and d.   □

Remark 1.11 The density property above says that, between any two real numbers is a rational number. Even more can be said. In fact **(i)** between every two irrational numbers is a rational number and **(ii)** between every two rational numbers is an irrational number.

In Appendix II at the end of the book we establish that the set of all decimal representations of numbers is uncountable. It follows that the set of all real numbers is uncountable. In fact the same proof shows that the set of all real numbers in the interval $(0,1)$, or in any nonempty open interval (c,d), is uncountable.

The set $\mathbb{R}$ of real numbers is uncountable (see Section A2.7 in Appendix II), yet the set $\mathbb{Q}$ of rational numbers is countable. It follows that the set $\mathbb{R} \setminus \mathbb{Q}$ of *irrational* numbers is uncountable. In particular, it is nonempty. Thus we may see with very little effort that there exist a great many real numbers which cannot be expressed as a quotient of integers. However, it can be quite difficult to see whether any particular real number (such as π or e or $\sqrt[5]{2}$) is irrational.

We conclude by recalling the "absolute value" notation:

Definition 1.12 Let x be a real number. We define

$$|x| = \begin{cases} x & \text{if} & x > 0 \\ 0 & \text{if} & x = 0 \\ -x & \text{if} & x < 0 \end{cases}$$

It is left as an exercise for you to verify the important *triangle inequality*:

$$|x + y| \le |x| + |y|.$$

[**Hint:** It is convenient to verify that the square of the left-hand side is less than or equal to the square of the right-hand side. See Exercise **7** below.]

APPENDIX: Construction of the Real Numbers

There are several techniques for constructing the real number system $\mathbb{R}$ from the rational numbers system $\mathbb{Q}$ (see Appendix I for a discussion of the origin of the rational numbers). We use the method of Dedekind (Julius W. R. Dedekind, 1831–1916) cuts because it uses a minimum of new ideas and is fairly brief.

The number system that we shall be constructing is an instance of a *field* (the complex numbers, in the next section, also form a field). The definition is as follows:

Definition 1.13 A set S is called a *field* if it is equipped with a binary operation (usually called addition and denoted "+") and a second binary operation (called multiplication and denoted "·") such that the following axioms are satisfied (Here A stands for "addition," M stands for "multiplication," and D stands for "distributive law."):

 A1. S is closed under addition: if $x, y \in S$ then $x + y \in S$.

A2. Addition is commutative: if $x, y \in S$ then $x + y = y + x$.

A3. Addition is associative: if $x, y, z \in S$ then $x + (y + z) = (x + y) + z$.

A4. There exists an element, called 0, in S which is an additive identity: if $x \in S$ then $0 + x = x$.

A5. Each element of S has an additive inverse: if $x \in S$ then there is an element $-x \in S$ such that $x + (-x) = 0$.

M1. S is closed under multiplication: if $x, y \in S$ then $x \cdot y \in S$.

M2. Multiplication is commutative: if $x, y \in S$ then $x \cdot y = y \cdot x$.

M3. Multiplication is associative: if $x, y, z \in S$ then $x \cdot (y \cdot z) = (x \cdot y) \cdot z$.

M4. There exists an element, called 1, which is a multiplicative identity: if $x \in S$ then $x \cdot 1 = x$.

M5. Each nonzero element of S has a multiplicative inverse: if $0 \neq x \in S$ then there is an element $x^{-1} \in S$ such that $x \cdot (x^{-1}) = 1$. The element x^{-1} is sometimes denoted $1/x$.

D1. Multiplication distributes over addition: if $x, y, z \in S$ then

$$x \cdot (y + z) = x \cdot y + x \cdot z.$$

Definition 1.14 A *cut* is a subset $\mathcal{C}$ of $\mathbb{Q}$ with the following properties:

- $\mathcal{C} \neq \emptyset$

- If $s \in \mathcal{C}$ and $t < s$ then $t \in \mathcal{C}$

- If $s \in \mathcal{C}$ then there is a $u \in \mathcal{C}$ such that $u > s$

- There is a rational number x such that $c < x$ for all $c \in \mathcal{C}$

You should think of a cut $\mathcal{C}$ as the set of all rational numbers to the left of some point in the real line. Since we have not constructed the real line yet, we cannot define a cut in that simple way; we have to make the construction more indirect. But if you consider the four properties of a cut, they describe a set that looks like a "rational halfline."

Notice that, if $\mathcal{C}$ is a cut and $s \notin \mathcal{C}$, then any rational $t > s$ is also not in $\mathcal{C}$. Also, if $r \in \mathcal{C}$ and $s \notin \mathcal{C}$ then it must be that $s > r$.

Definition 1.15 If $\mathcal{C}$ and $\mathcal{D}$ are cuts then we say that $\mathcal{C} < \mathcal{D}$ provided that $\mathcal{C}$ is a subset of $\mathcal{D}$ but $\mathcal{C} \neq \mathcal{D}$.

Check for yourself that "<" is an ordering on the set of all cuts.

Now we introduce operations of addition and multiplication which will turn the set of all cuts into a field.

Definition 1.16 If $\mathcal{C}$ and $\mathcal{D}$ are cuts then we define

$$\mathcal{C} + \mathcal{D} = \{c + d : c \in \mathcal{C}, d \in \mathcal{D}\}.$$

We define the cut $\widehat{0}$ to be the set of all negative rationals.

The cut $\widehat{0}$ will play the role of the additive identity. We are now required to check that field axioms **A1**–**A5** hold.

For **A1**, we need to see that $\mathcal{C} + \mathcal{D}$ is a cut. Obviously $\mathcal{C} + \mathcal{D}$ is not empty. If s is an element of $\mathcal{C} + \mathcal{D}$ and t is a rational number less than s, write $s = c + d$, where $c \in \mathcal{C}$ and $d \in \mathcal{D}$. Then $t - c < s - c = d \in \mathcal{D}$ so $t - c \in \mathcal{D}$; and $c \in \mathcal{C}$. Hence $t = c + (t - c) \in \mathcal{C} + \mathcal{D}$. A similar argument shows that there is an $r > s$ such that $r \in \mathcal{C} + \mathcal{D}$. Finally, if x is a rational upper bound for $\mathcal{C}$ and y is a rational upper bound for $\mathcal{D}$, then $x + y$ is a rational upper bound for $\mathcal{C} + \mathcal{D}$. We conclude that $\mathcal{C} + \mathcal{D}$ is a cut.

Since addition of rational numbers is commutative, it follows immediately that addition of cuts is commutative. Associativity follows in a similar fashion.

Now we show that if $\mathcal{C}$ is a cut then $\mathcal{C} + \widehat{0} = \mathcal{C}$. For if $c \in \mathcal{C}$ and $z \in \widehat{0}$ then $c + z < c + 0 = c$ hence $\mathcal{C} + \widehat{0} \subseteq \mathcal{C}$. Also, if $c' \in \mathcal{C}$ then choose a $d' \in \mathcal{C}$ such that $c' < d'$. Then $c' - d' < 0$ so $c' - d' \in \widehat{0}$. And $c' = d' + (c' - d')$. Hence $\mathcal{C} \subseteq \mathcal{C} + \widehat{0}$. We conclude that $\mathcal{C} + \widehat{0} = \mathcal{C}$.

Finally, for Axiom **A5**, we let $\mathcal{C}$ be a cut and set $-\mathcal{C}$ to be equal to $\{d \in \mathbb{Q} : c + d < 0 \text{ for all } c \in \mathcal{C}\}$. If x is a rational upper bound for $\mathcal{C}$ and $c \in \mathcal{C}$ then $-x \in -\mathcal{C}$ so $-\mathcal{C}$ is not empty. By its very definition, $\mathcal{C} + (-\mathcal{C}) \subseteq \widehat{0}$. Further, if $z \in \widehat{0}$ and $c \in \mathcal{C}$ we set $c' = z - c$. Then $c' \in -\mathcal{C}$ and $z = c + c'$. Hence $\widehat{0} \subseteq \mathcal{C} + (-\mathcal{C})$. We conclude that $\mathcal{C} + (-\mathcal{C}) = \widehat{0}$.

Having verified the axioms for addition, we turn now to multiplication.

Definition 1.17 If $\mathcal{C}$ and $\mathcal{D}$ are cuts then we define the product $\mathcal{C} \cdot \mathcal{D}$ as follows:

- If $\mathcal{C}, \mathcal{D} > \widehat{0}$ then $\mathcal{C} \cdot \mathcal{D} = \{q \in \mathbb{Q} : q < c \cdot d \text{ for some } c \in \mathcal{C}, d \in \mathcal{D} \text{ with } c > 0, d > 0\}$

- If $\mathcal{C} > \widehat{0}, \mathcal{D} < \widehat{0}$ then $\mathcal{C} \cdot \mathcal{D} = -(\mathcal{C} \cdot (-\mathcal{D}))$

- If $\mathcal{C} < \widehat{0}, \mathcal{D} > \widehat{0}$ then $\mathcal{C} \cdot \mathcal{D} = -((-\mathcal{C}) \cdot \mathcal{D})$

- If $\mathcal{C}, \mathcal{D} < \widehat{0}$ then $\mathcal{C} \cdot \mathcal{D} = (-\mathcal{C}) \cdot (-\mathcal{D})$

- If either $\mathcal{C} = \widehat{0}$ or $\mathcal{D} = \widehat{0}$ then $\mathcal{C} \cdot \mathcal{D} = \widehat{0}$.

Notice that, for convenience, we have defined multiplication of negative numbers just as we did in high school. The reason is that the definition that we use for the product of two positive numbers cannot work when one of the two factors is negative (exercise).

It is now a routine exercise to verify that the set of all cuts, with this definition of multiplication, satisfies field axioms **M1**–**M5**. The proofs follow those for **A1**–**A5** rather closely.

For the distributive property, one first checks the case when all the cuts are positive, reducing it to the distributive property for the rationals. Then one handles negative cuts on a case by case basis.

We now know that the collection of all cuts forms an ordered field. Denote this field by the symbol $\mathbb{R}$. We next verify the crucial property of $\mathbb{R}$ that sets it apart from $\mathbb{Q}$:

Theorem 1.18 *The ordered field $\mathbb{R}$ satisfies the least upper bound property.*

Proof: Let S be a subset of $\mathbb{R}$ which is bounded above. Define

$$\mathcal{S}^* = \bigcup_{\mathcal{C} \in S} \mathcal{C}.$$

Then $\mathcal{S}^*$ is clearly nonempty, and it is therefore a cut since it is a union of cuts. It is also clearly an upper bound for S since it contains each element of S. It remains to check that $\mathcal{S}^*$ is the least upper bound for S.

In fact if $\mathcal{T} < \mathcal{S}^*$ then $\mathcal{T} \subseteq \mathcal{S}^*$ and there is a rational number q in $\mathcal{S}^* \setminus \mathcal{T}$. But, by the definition of $\mathcal{S}^*$, it must be that $q \in \mathcal{C}$ for some $\mathcal{C} \in S$. So $\mathcal{C} > \mathcal{T}$, and $\mathcal{T}$ cannot be an upper bound for S. Therefore $\mathcal{S}^*$ is the least upper bound for S, as desired.   □

We have shown that $\mathbb{R}$ is an ordered field which satisfies the least upper bound property. It remains to show that $\mathbb{R}$ contains (a copy of) $\mathbb{Q}$ in a natural way. In fact, if $q \in \mathbb{Q}$ we associate to it the element $\varphi(q) = \mathcal{C}_q \equiv \{x \in \mathbb{Q} : x < q\}$. Then $\mathcal{C}_q$ is obviously a cut. It is also routine to check that

$$\varphi(q + q') = \varphi(q) + \varphi(q') \quad \text{and} \quad \varphi(q \cdot q') = \varphi(q) \cdot \varphi(q').$$

Therefore we see that φ represents $\mathbb{Q}$ as a subfield of $\mathbb{R}$.

Exercises

1. Give an example of a set of real numbers that contains its least upper bound but not its greatest lower bound. Give an example of a set that contains its greatest lower bound but not its least upper bound.

2. Give an example of a set of real numbers that does *not* have a least upper bound. Give an example of a set of real numbers that does *not* have a greatest lower bound.

3. Let A be a set of real numbers that is bounded above and set $\alpha = \sup A$. Let $B = \{-a : a \in A\}$. Prove that $\inf B = -\alpha$. Prove the same result with the roles of infimum and supremum reversed.

4. What is the least upper bound of the set

$$S = \{x : x^2 < 2\} ?$$

Explain why this question has a sensible answer in the real number system but not in the rational number system.

5. Prove that the least upper bound and greatest lower bound for a set of real numbers is each unique.

6. Consider the unit circle C (the circle with center the origin in the plane and radius 1). Let

$$S = \{\alpha : 2\alpha < \text{ (the circumference of } C)\}.$$

Show that S is bounded above. Let p be the least upper bound of S. Say explicitly what the number p is. This exercise works in the real number system, but not in the rational number system. Why?

7. Prove the triangle inequality.

8. Let $\emptyset$ be the empty set—the set with no elements. Prove that $\sup \emptyset = -\infty$ and $\inf \emptyset = +\infty$.

9. Prove that addition of the real numbers (as constructed in the Appendix) is commutative. Now prove that it is associative.

10. Complete the calculation in the proof of Theorem 1.7.

11. Describe a countable set of nonrational real numbers between 0 and 1.

* 12. Let f be a function with domain the reals and range the reals. Assume that f has a local minimum at each point x in its domain. (This means that, for each $x \in \mathbb{R}$, there is an $\epsilon = \epsilon_x > 0$ such that, whenever $\mid x - t \mid < \epsilon$ then $f(x) \leq f(t)$.) *Do not assume that f is differentiable, or continuous, or anything nice like that.* Prove that the image of f is countable. (**Hint:** When I solved this problem as a student my solution was ten pages long; however, there is a one-line solution due to Michael Spivak.)

* 13. Let λ be a positive irrational real number. If n is a positive integer, choose by the Archimedean Property an integer k such that $k\lambda \leq n < (k+1)\lambda$. Let $\varphi(n) = n - k\lambda$. Prove that the set of all $\varphi(n)$, $n > 0$, is dense in the interval $[0, \lambda]$. (**Hint:** Examine the proof of the density of the rationals in the reals.)

* 14. Let n be a natural number and x a positive real number. Prove that there is a positive real number y such that $y^n = x$. Is y unique?

1.2 The Complex Numbers

When we first learn about the complex numbers, the most troublesome point is the very beginning: "Let's pretend that the number -1 has a square root. Call it i." What gives us the right to "pretend" in this fashion? The answer is

that we have no such right.[1] If -1 has a square root, then we should be able to construct a number system in which that is the case. That is what we shall do in this section.

Definition 1.19 The system of *complex numbers*, denoted by the symbol $\mathbb{C}$, consists of all ordered pairs (a, b) of real numbers. We add two complex numbers (a, b) and $(\widetilde{a}, \widetilde{b})$ by the formula

$$(a, b) + (\widetilde{a}, \widetilde{b}) = (a + \widetilde{a}, b + \widetilde{b}).$$

We multiply two complex numbers by the formula

$$(a, b) \cdot (\widetilde{a}, \widetilde{b}) = (a \cdot \widetilde{a} - b \cdot \widetilde{b}, a \cdot \widetilde{b} + \widetilde{a} \cdot b).$$

Remark 1.20 If you are puzzled by this definition of multiplication, do not worry. In a few moments you will see that it gives rise to the notion of multiplication of complex numbers that you are accustomed to. Perhaps more importantly, a naive rule for multiplication like $(a, b) \cdot (\widetilde{a}, \widetilde{b}) = (a\widetilde{a}, b\widetilde{b})$ gives rise to nonsense like $(1, 0) \cdot (0, 1) = (0, 0)$. It is really necessary for us to use the initially counterintuitive definition of multiplication that is presented here.

EXAMPLE 1.21 Let $z = (3, -2)$ and $w = (4, 7)$ be two complex numbers. Then

$$z + w = (3, -2) + (4, 7) = (3 + 4, -2 + 7) = (7, 5).$$

Also

$$z \cdot w = (3, -2) \cdot (4, 7) = (3 \cdot 4 - (-2) \cdot 7, 3 \cdot 7 + 4 \cdot (-2)) = (26, 13). \qquad \square$$

As usual, we ought to check that addition and multiplication are commutative, associative, that multiplication distributes over addition, and so forth. We shall leave these tasks to the exercises. Instead we develop some of the crucial, and more interesting, properties of our new number system.

Theorem 1.22 *The following properties hold for the number system $\mathbb{C}$.*

(a) *The number $1 \equiv (1, 0)$ is the multiplicative identity: $1 \cdot z = z \cdot 1 = z$ for any $z \in \mathbb{C}$.*

(b) *The number $0 \equiv (0, 0)$ is the additive identity: $0 + z = z + 0 = z$ for any $z \in \mathbb{C}$.*

---

[1] The complex numbers were initially developed so that we would have a number system in which all polynomial equations are solvable. One of the reasons, historically, that mathematicians had trouble accepting the complex numbers is that they did not believe that they really existed—they were just made up. This is, in part, how they came to be called "imaginary" and "complex." Mathematicians had similar trouble accepting negative numbers; for a time, negative numbers were called "forbidden."

(c) *Each complex number $z = (x, y)$ has an additive inverse $-z = (-x, -y)$: it holds that $z + (-z) = (-z) + z = 0$.*

(d) *The number $i \equiv (0, 1)$ satisfies $i \cdot i = -1$; in other words, i is a square root of -1.*

Proof: These are direct calculations, but it is important for us to work out these facts.

First, let $z = (x, y)$ be any complex number. Then

$$1 \cdot z = (1, 0) \cdot (x, y) = (1 \cdot x - 0 \cdot y, 1 \cdot y + x \cdot 0) = (x, y) = z \,.$$

This proves the first assertion.

For the second, we have

$$0 + z = (0, 0) + (x, y) = (0 + x, 0 + y) = (x, y) = z \,.$$

With z as above, set $-z = (-x, -y)$. Then

$$z + (-z) = (x, y) + (-x, -y) = (x + (-x), y + (-y)) = (0, 0) = 0 \,.$$

Finally, we calculate

$$i \cdot i = (0, 1) \cdot (0, 1) = (0 \cdot 0 - 1 \cdot 1, 0 \cdot 1 + 0 \cdot 1) = (-1, 0) = -1 \,.$$

Thus, as asserted, i is a square root of -1. $\qquad\square$

Proposition 1.23 *If $z \in \mathbb{C}, z \neq 0$, then there is a complex number w such that $z \cdot w = 1$.*

Proof: You might be thinking, "Well, of course $w = 1/z$. But this is nonsense. The expression $1/z$ is *not* written in the form of a complex number!

Write $z = (x, y)$ and set

$$w = \left(\frac{x}{x^2 + y^2}, \frac{-y}{x^2 + y^2} \right) \,.$$

Since $z \neq 0$, $x^2 + y^2 \neq 0$, so this definition makes sense. Then it is straightforward to verify that $z \cdot w = 1$:

$$z \cdot w = (x, y) \cdot \left(\frac{x}{x^2 + y^2}, \frac{-y}{x^2 + y^2} \right)$$

$$= \left(x \cdot \frac{x}{x^2 + y^2} - y \cdot \frac{-y}{x^2 + y^2}, x \cdot \frac{-y}{x^2 + y^2} + \frac{x}{x^2 + y^2} \cdot y \right)$$

$$= \left(\frac{x^2 + y^2}{x^2 + y^2}, \frac{-yx + xy}{x^2 + y^2} \right)$$

$$= (1, 0)$$

$$= 1 \,.$$

$\qquad\square$

EXAMPLE 1.24 Consider the complex number $z = 2 + 3i$. According to the proposition,

$$w = \left(\frac{2}{2^2 + 3^2}, \frac{-3}{2^2 + 3^2} \right) = \left(\frac{2}{13}, \frac{-3}{13} \right)$$

will be the multiplicative inverse of z. And, indeed,

$$z \cdot w = (2, 3) \cdot \left(\frac{2}{13}, \frac{-3}{13} \right) = \left(\frac{4}{13} + \frac{9}{13}, \frac{-6}{13} + \frac{6}{13} \right) = \left(\frac{13}{13}, 0 \right) = (1, 0) = 1 \,. \ \Box$$

Of course we interpret the quotient z/w of complex numbers to mean $z \cdot (1/w)$. This will be a new complex number.

Thus every nonzero complex number has a multiplicative inverse. The other field axioms for $\mathbb{C}$ are easy to check. We conclude that the number system $\mathbb{C}$ forms a field. You will prove in the exercises that it is not possible to order this field. If α is a real number then we associate α with the complex number $(\alpha, 0)$. Thus we have the natural "embedding"

$$\mathbb{R} \ni \alpha \longmapsto (\alpha, 0) \in \mathbb{C} \,.$$

In this way, we can think of the real numbers as a *subset* of the complex numbers. In fact, the real field $\mathbb{R}$ is a *subfield* of the complex field $\mathbb{C}$. This means that if $\alpha, \beta \in \mathbb{R}$ and $(\alpha, 0), (\beta, 0)$ are the corresponding elements in $\mathbb{C}$ then $\alpha + \beta$ corresponds to $(\alpha + \beta, 0)$ and $\alpha \cdot \beta$ corresponds to $(\alpha \cdot \beta, 0)$. These assertions are explored more thoroughly in the exercises.

With the remarks in the preceding paragraph we can sometimes ignore the distinction between the real numbers and the complex numbers. For example, we can write

$$5 \cdot i$$

and understand that it means $(5, 0) \cdot (0, 1) = (0, 5)$. Likewise, the expression

$$5 \cdot 1$$

can be interpreted as $5 \cdot 1 = 5$ or as $(5, 0) \cdot (1, 0) = (5, 0)$ without any danger of ambiguity.

Theorem 1.25 *Every complex number can be written in the form $a + b \cdot i$, where a and b are real numbers. In fact, if $z = (x, y) \in \mathbb{C}$ then*

$$z = x + y \cdot i \,.$$

Proof: With the identification of real numbers as a subfield of the complex numbers, we have that

$$x + y \cdot i = (x, 0) + (y, 0) \cdot (0, 1) = (x, 0) + (0, y) = (x, y) = z$$

as claimed. $\Box$

Now that we have constructed the complex number field, we will adhere to the usual custom of writing complex numbers as $z = a + b \cdot i$ or, more simply, $a + bi$. We call a the *real part* of z, denoted by $\operatorname{Re} z$, and b the *imaginary part* of z, denoted $\operatorname{Im} z$. We have

$$(a + bi) + (\tilde{a} + \tilde{b}i) = (a + \tilde{a}) + (b + \tilde{b})i$$

and

$$(a + bi) \cdot (\tilde{a} + \tilde{b}i) = (a \cdot \tilde{a} - b \cdot \tilde{b}) + (a \cdot \tilde{b} + \tilde{a} \cdot b)i \,.$$

EXAMPLE 1.26 Let $z = 3 - 7i$ and $w = 4 + 6i$. Then

$$z + w = (3 - 7i) + (4 + 6i) = 7 - i \,,$$

$$\begin{aligned}
z \cdot w &= (3 - 7i) \cdot (4 + 6i) \\
&= (3 \cdot 4 - (-7) \cdot 6) + (3 \cdot 6 + (-7) \cdot 4)i \\
&= 54 - 10i \,.
\end{aligned}$$ $\qquad\square$

If $z = a + bi$ is a complex number then we define its *complex conjugate* to be the number $\overline{z} = a - bi$. We record some elementary facts about the complex conjugate:

Proposition 1.27 *If z, w are complex numbers then*

(1) $\overline{z + w} = \overline{z} + \overline{w}$;

(2) $\overline{z \cdot w} = \overline{z} \cdot \overline{w}$;

(3) $z + \overline{z} = 2 \cdot \operatorname{Re} z$;

(4) $z - \overline{z} = 2 \cdot i \cdot \operatorname{Im} z$;

(5) $z \cdot \overline{z} \geq 0$, *with equality holding if and only if $z = 0$.*

Proof: Write $z = a + bi, w = c + di$. Then

$$\begin{aligned}
\overline{z + w} &= \overline{(a + c) + (b + d)i} \\
&= (a + c) - (b + d)i \\
&= (a - bi) + (c - di) \\
&= \overline{z} + \overline{w}.
\end{aligned}$$

This proves **(1)**. Assertions **(2)**, **(3)**, **(4)** are proved similarly.
For **(5)**, notice that

$$z \cdot \overline{z} = (a + bi) \cdot (a - bi) = a^2 + b^2 \geq 0.$$

Clearly equality holds if and only if $a = b = 0$. $\qquad\square$

EXAMPLE 1.28 Let $z = -2 + 4i$ and $w = 5 - 3i$. Then

$$\overline{z} = -2 - 4i.$$

Also

$$\overline{z \cdot w} = \overline{(-2 + 4i)(5 - 3i)} = \overline{2 + 26i} = 2 - 26i$$

while

$$\overline{z} \cdot \overline{w} = (-2 - 4i) \cdot (5 + 3i) = (-10 + 12) + (-6 - 20)i = 2 - 26i. \qquad \square$$

The expression $|z|$ is defined to be the nonnegative square root of $z \cdot \overline{z}$:

$$|z| = +\sqrt{z \cdot \overline{z}} = \sqrt{x^2 + y^2}$$

when $z = x + iy$. It is called the *modulus* of z and plays the same role for the complex field that absolute value plays for the real field. It is the distance of z to the origin. The modulus has the following properties.

Proposition 1.29 *If* $z, w \in \mathbb{C}$ *then*

(1) $|z| = |\overline{z}|$;

(2) $|z \cdot w| = |z| \cdot |w|$;

(3) $|\mathrm{Re}\, z| \leq |z|$, $|\mathrm{Im}\, z| \leq |z|$;

(4) $|z + w| \leq |z| + |w|$;

Proof: Write $z = a + bi, w = c + di$. Then **(1)**, **(2)**, **(3)** are immediate. For **(4)** we calculate that

$$
\begin{aligned}
|z + w|^2 &= (z + w) \cdot (\overline{z + w}) \\
&= z \cdot \overline{z} + z \cdot \overline{w} + w \cdot \overline{z} + w \cdot \overline{w} \\
&= |z|^2 + 2\mathrm{Re}\,(z \cdot \overline{w}) + |w|^2 \\
&\leq |z|^2 + 2|z \cdot \overline{w}| + |w|^2 \\
&= |z|^2 + 2|z| \cdot |w| + |w|^2 \\
&= (|z| + |w|)^2.
\end{aligned}
$$

Taking square roots proves **(4)**. $\square$

EXAMPLE 1.30 Let $z = 3 + 4i$ and $w = -5 + 2i$. Then

$$|z| = \sqrt{3^2 + 4^2} = \sqrt{25} = 5 \qquad \text{and} \qquad |w| = \sqrt{(-5)^2 + 2^2} = \sqrt{29}.$$

Also

$$|z \cdot w| = |(3 + 4i)(-5 + 2i)| = |-23 - 14i| = \sqrt{23^2 + 14^2} = \sqrt{725} = 5\sqrt{29}$$

while

$$|z| \cdot |w| = 5 \cdot \sqrt{29}. \qquad \square$$
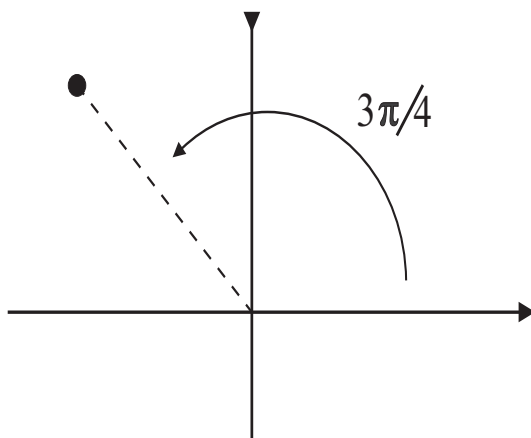
Figure 1.1: The polar form of $-1 + i$.

Observe that, if z is real, then $z = a + 0i$ and the modulus of z equals the absolute value of a. Likewise, if $z = 0 + bi$ is pure imaginary, then the modulus of z equals the absolute value of b. In particular, the fourth part of the proposition reduces, in the real case, to the triangle inequality

$$|a + b| \leq |a| + |b| \,.$$

If z is any nonzero complex number, then let $r = |z|$. Now define $\xi = z/r$. We see that ξ is a complex number of modulus 1. Thus ξ lies on the unit circle, so it subtends an angle θ with the positive x-axis. Then $\xi = \cos\theta + i\sin\theta$. It is shown in Section 9.3 that

$$e^{i\theta} = \xi = \cos\theta + i\sin\theta \,.$$

[**Hint:** You may verify this formula for yourself by writing out the power series for the exponential and writing out the power series for cosine and sine.] We often call

$$z = re^{i\theta}$$

the *polar form* of z.

EXAMPLE 1.31 Let us find all cube roots of the complex number $z = -1 + i$. Using the notation of the preceding paragraph, we see that $r = \sqrt{(-1)^2 + 1^2} = \sqrt{2}$. Thus $\xi = z/r = -1/\sqrt{2} + (1/\sqrt{2})i$. Examining Figure 1.1, we see that $\theta = 3\pi/4$. We have learned then that

$$z = -1 + i = \sqrt{2}e^{i3\pi/4} \,.$$

For the first cube root w_1 of z, we write $w_1 = se^{i\psi}$, and we solve for s and ψ. We know that

$$(w_1)^3 = z$$

so
$$\left(se^{i\psi}\right)^3 = \sqrt{2}e^{i3\pi/4}$$

or
$$s^3 e^{i3\psi} = \sqrt{2}e^{i3\pi/4}\,.$$

It is natural then to conclude that
$$s^3 = \sqrt{2}$$

and
$$3\psi = 3\pi/4\,.$$

We conclude that $s = 2^{1/6}$ and $\psi = \pi/4$. We have found that
$$w_1 = 2^{1/6}e^{i\pi/4}$$

is a cube root of z. But this is not the only cube root! There are three cube roots in total.

We next notice that z can also be written
$$z = \sqrt{2}e^{i((3\pi/4)+2\pi)}\,.$$

[Observe that there is some ambiguity built into the polar form of a complex number, just as there is ambiguity in the polar coordinates that you learned about in calculus. The reason is that the cosine and sine functions are 2π-periodic.]

Now we repeat the calculation above with this new form for the complex number z. We know that
$$(w_2)^3 = z$$

so
$$\left(se^{i\psi}\right)^3 = \sqrt{2}e^{i((3\pi/4)+2\pi)}$$

or
$$s^3 e^{i3\psi} = \sqrt{2}e^{i11\pi/4}\,.$$

It is natural then to conclude that
$$s^3 = \sqrt{2}$$

and
$$3\psi = 11\pi/4\,.$$

We conclude that $s = 2^{1/6}$ and $\psi = 11\pi/12$. We have found that
$$w_2 = 2^{1/6}e^{i11\pi/12}$$

is a cube root of z.

Let us do the calculation one more time with z now written as
$$z = \sqrt{2}e^{i((3\pi/4)+4\pi)}$$

(again we exploit the periodicity of sine and cosine). We know that

$$(w_3)^3 = z$$

so

$$\left(se^{i\psi}\right)^3 = \sqrt{2}e^{i((3\pi/4)+4\pi)}$$

or

$$s^3 e^{i3\psi} = \sqrt{2}e^{i19\pi/4}.$$

It is natural then to conclude that

$$s^3 = \sqrt{2}$$

and

$$3\psi = 19\pi/4.$$

We conclude that $s = 2^{1/6}$ and $\psi = 19\pi/12$. We have found that

$$w_3 = 2^{1/6}e^{i19\pi/12}$$

is a cube root of z.

There is no sense to repeat these calculations any further. It is true that $z = \sqrt{2}e^{i(3\pi/4+6\pi)}$. But performing our calculations for this form of z would simply cause us to rediscover w_1. We have found three cube roots of z, and that is the end of the calculation. □

We conclude this discussion by recording the most important basic fact about the complex numbers. Carl Friedrich Gauss gave five proofs of this theorem (the Fundamental Theorem of Algebra) in his doctoral dissertation:

Theorem 1.32 *Let $p(z)$ be any polynomial of degree at least 1. Then p has a root $\alpha \in \mathbb{C}$ such that $p(\alpha) = 0$.*

Using a little algebra, one can in fact show that a polynomial of degree k has k roots (counting multiplicity).

Exercises

1. Show that, if z is a nonzero complex number, then its multiplicative inverse is given by
$$w = \frac{\overline{z}}{|z|^2}.$$

2. Refer to Exercise 1. If $z, w \in \mathbb{C}$ then prove that $\overline{z/w} = \overline{z}/\overline{w}$.

3. Find all cube roots of the complex number $1 + i$.

4. Taking the commutative, associative, and distributive laws of addition and multiplication for the real number system for granted, establish these laws for the complex numbers.

5. Consider the function $\phi : \mathbb{R} \to \mathbb{C}$ given by $\phi(x) = x + i \cdot 0$. Prove that ϕ respects addition and multiplication in the sense that $\phi(x + x') = \phi(x) + \phi(x')$ and $\phi(x \cdot x') = \phi(x) \cdot \phi(x')$.

6. Prove that the field of complex numbers cannot be made into an *ordered* field. (**Hint:** Since $i \neq 0$ then either $i > 0$ or $i < 0$. Both lead to a contradiction.)

7. Prove that the complex roots of a polynomial with real coefficients occur in complex conjugate pairs.

8. Calculate the square roots of i.

9. Prove that the set of all complex numbers is uncountable.

10. Prove that any nonzero complex number z has kth roots $r_1, r_2, \ldots, r_k$. That is, prove that there are k of them.

11. In the complex plane, draw a picture of

$$S = \{ z \in \mathbb{C} : |z - 1| + |z + 1| = 2 \} \, .$$

12. Refer to Exercise 9. Show that the kth roots of z all lie on a circle centered at the origin, and that they are equally spaced.

13. Find all the cube roots of $1 + i$.

14. Find all the square roots of $-1 - i$.

15. Prove that the set of all complex numbers with rational real part is uncountable.

16. Prove that the set of all complex numbers with both real and imaginary parts rational is countable.

17. Prove that the set $\{ z \in \mathbb{C} : |z| = 1 \}$ is uncountable.

* 18. In the complex plane, draw a picture of

$$T = \{ z \in \mathbb{C} : |z + \overline{z}| - |z - \overline{z}| = 2 \} \, .$$

* 19. Use the Fundamental Theorem of Algebra to prove that any polynomial of degree k has k (not necessarily distinct) roots. [**Hint:** Use the Euclidean algorithm.]

Chapter 2

Sequences

2.1 Convergence of Sequences

A *sequence* of real numbers is a function $\varphi : \mathbb{N} \to \mathbb{R}$. We often write the sequence as $\varphi(1), \varphi(2), \ldots$ or, more simply, as $\varphi_1, \varphi_2, \ldots$. A sequence of complex numbers is defined similarly, with $\mathbb{R}$ replaced by $\mathbb{C}$.

EXAMPLE 2.1 The function $\varphi(j) = 1/j$ is a sequence of real numbers. We will often write such a sequence as $\varphi_j = 1/j$ or as $\{1, 1/2, 1/3, \ldots\}$ or as $\{1/j\}_{j=1}^{\infty}$.

The function $\psi(j) = \cos j + i \sin j$ is a sequence of complex numbers.

Do not be misled into thinking that a sequence must form a pattern, or be given by a formula. Obviously the ones which are given by formulas are easy to write down, but they are not typical. For example, the coefficients in the decimal expansion of $\pi, \{3, 1, 4, 1, 5, 9, 2, 6, 5, \ldots\}$, fit our definition of sequence—but they are not given by any obvious pattern. □

The most important question about a sequence is whether it converges. We define this notion as follows.

Definition 2.2 A sequence $\{a_j\}$ of real (resp. complex) numbers is said to *converge* to a real (resp. complex) number α if, for each $\epsilon > 0$, there is an integer $N > 0$ such that, if $j > N$, then $|a_j - \alpha| < \epsilon$. We call α the *limit* of the sequence $\{a_j\}$. We write $\lim_{j \to \infty} a_j = \alpha$. We also sometimes write $a_j \to \alpha$.

If a sequence $\{a_j\}$ does not converge then we frequently say that it *diverges*.

EXAMPLE 2.3 Let $a_j = 1/j, j = 1, 2, \ldots$. Then the sequence converges to 0. For let $\epsilon > 0$. Choose N to be the next integer after $1/\epsilon$ (we use here the Archimedean principle). If $j > N$ then

$$|a_j - 0| = |a_j| = \frac{1}{j} < \frac{1}{N} < \epsilon,$$

proving the claim.

Let $b_j = (-1)^j, j = 1, 2, \ldots$. Then the sequence *does not converge*. To prove this assertion, suppose to the contrary that it does. Suppose that the sequence converges to a number α. Let $\epsilon = 1/2$. By definition of convergence, there is an integer $N > 0$ such that, if $j > N$, then $|b_j - \alpha| < \epsilon = 1/2$. For such j we have

$$|b_j - b_{j+1}| = |(b_j - \alpha) + (\alpha - b_{j+1})| \leq |b_j - \alpha| + |\alpha - b_{j+1}|$$

(by the triangle inequality—see the end of Section 1.1). But this last is

$$< \epsilon + \epsilon = 1 \, .$$

On the other hand,

$$|b_j - b_{j+1}| = \left|(-1)^j - (-1)^{j+1}\right| = 2 \, .$$

The last two lines yield that $2 < 1$, a clear contradiction. So the sequence $\{b_j\}$ has no limit.   □

We begin with a few intuitively appealing properties of convergent sequences which will be needed later. First, a definition.

Definition 2.4 A sequence a_j is said to be *bounded* if there is a number $M > 0$ such that $|a_j| \leq M$ for every j.

Now we have

Proposition 2.5 *Let $\{a_j\}$ be a convergent sequence. Then we have:*

- *The limit of the sequence is unique.*

- *The sequence is bounded.*

Proof: Suppose that the sequence has two limits α and $\widetilde{\alpha}$. Let $\epsilon > 0$. Then there is an integer $N > 0$ such that for $j > N$ we have the inequality $|a_j - \alpha| < \epsilon/2$. Likewise, there is an integer $\widetilde{N} > 0$ such that for $j > \widetilde{N}$ we have $|a_j - \widetilde{\alpha}| < \epsilon/2$. Let $N_0 = \max\{N, \widetilde{N}\}$. Then, for $j > N_0$, we have

$$|\alpha - \widetilde{\alpha}| = |(\alpha - a_j) + (a_j - \widetilde{\alpha})| \leq |\alpha - a_j| + |a_j - \widetilde{\alpha}| < \epsilon/2 + \epsilon/2 = \epsilon \, .$$

Since this inequality holds for any $\epsilon > 0$ we have that $\alpha = \widetilde{\alpha}$.

Next, with α the limit of the sequence and $\epsilon = 1$, we choose an integer $N > 0$ such that $j > N$ implies that $|a_j - \alpha| < \epsilon = 1$. For such j we have that

$$|a_j| = |(a_j - \alpha) + \alpha| \leq |a_j - \alpha| + |\alpha| < 1 + |\alpha| \equiv P \, .$$

Let $Q = \max\{|a_1|, |a_2|, \ldots, |a_N|\}$. If j is any natural number then either $1 \leq j \leq N$ (in which case $|a_j| \leq Q$) or else $j > N$ (in which case $|a_j| \leq P$). Set $M = \max\{P, Q\}$. Then $|a_j| \leq M$ for all j, as desired. So the sequence is bounded.   □

The next proposition records some elementary properties of limits of sequences.

Proposition 2.6 *Let* $\{a_j\}$ *be a sequence of real or complex numbers with limit* α *and* $\{b_j\}$ *be a sequence of real or complex numbers with limit* β. *Then we have:*

(1) *If* c *is a constant then the sequence* $\{c \cdot a_j\}$ *converges to* $c \cdot \alpha$;

(2) *The sequence* $\{a_j + b_j\}$ *converges to* $\alpha + \beta$;

(3) *The sequence* $a_j \cdot b_j$ *converges to* $\alpha \cdot \beta$;

(4) *If* $b_j \neq 0$ *for all* j *and* $\beta \neq 0$ *then the sequence* a_j/b_j *converges to* α/β.

Proof: For the first part, we may assume that $c \neq 0$ (for when $c = 0$ there is nothing to prove). Let $\epsilon > 0$. Choose an integer $N > 0$ such that for $j > N$ it holds that

$$|a_j - \alpha| < \frac{\epsilon}{|c|}\,.$$

For such j we have that

$$|c \cdot a_j - c \cdot \alpha| = |c| \cdot |a_j - \alpha| < |c| \cdot \frac{\epsilon}{|c|} = \epsilon\,.$$

This proves the first assertion.

The proof of the second assertion is similar, and we leave it as an exercise.

For the third assertion, notice that the sequence $\{a_j\}$ is bounded (by the second part of Proposition 2.5): say that $|a_j| \leq M$ for every j. Let $\epsilon > 0$. Choose an integer $N > 0$ so that $|a_j - \alpha| < \epsilon/(2M + 2|\beta|)$ when $j > N$. Also choose an integer $\widetilde{N} > 0$ such that $|b_j - \beta| < \epsilon/(2M + 2|\beta|)$ when $j > \widetilde{N}$. Then, for $j > \max\{N, \widetilde{N}\}$, we have that

$$
\begin{aligned}
|a_j b_j - \alpha\beta| &= |a_j(b_j - \beta) + \beta(a_j - \alpha)| \\
&\leq |a_j(b_j - \beta)| + |\beta(a_j - \alpha)| \\
&< M \cdot \frac{\epsilon}{2M + 2|\beta|} + |\beta| \cdot \frac{\epsilon}{2M + 2|\beta|} \\
&\leq \frac{\epsilon}{2} + \frac{\epsilon}{2} \\
&= \epsilon.
\end{aligned}
$$

So the sequence $\{a_j b_j\}$ converges to $\alpha\beta$.

Part **(4)** is proved in a similar fashion and we leave the details as an exercise. $\square$

Remark 2.7 You were probably puzzled by the choice of N and $\widetilde{N}$ in the proof of part **(3)** of Proposition 2.6—where did the number $\epsilon/(2M+2|\beta|)$ come from? The answer of course becomes obvious when we read on further in the proof. So the lesson here is that a proof is constructed backward: you look to the end of the proof to see what you need to specify earlier on. Skill in these matters can come only with practice.

EXAMPLE 2.8 Let $a_j = \sin(1/j)/(1/j)$ and $b_j = j^2/(2j^2+j)$. Then $a_j \to 1$ and $b_j \to 1/2$ as $j \to \infty$. Let us say a few words about why this is true.

You learned in your calculus class that

$$\lim_{x \to 0} \frac{\sin x}{x} = 1\,.$$

Letting $x = 1/j$ and $j \to \infty$ then yields that

$$\lim_{j \to \infty} \frac{\sin(1/j)}{1/j} = 1\,.$$

For the second limit, write

$$b_j = \frac{j^2}{2j^2 + j} = \frac{j^2/j^2}{(2j^2 + 1)/j^2} = \frac{1}{2 + 1/j^2}\,.$$

Now it is evident that

$$\lim_{j \to \infty} b_j = \frac{1}{2}\,.$$

From the above we may conclude, using Proposition 2.6, that

$$\lim_{j \to \infty} 5a_j = 5\,,$$

$$\lim_{j \to \infty} (a_j + b_j) = 1 + \frac{1}{2} = \frac{3}{2}\,,$$

$$\lim_{j \to \infty} a_j \cdot b_j = 1 \cdot \frac{1}{2} = \frac{1}{2}$$

and

$$\lim_{j \to \infty} \frac{a_j}{b_j} = \frac{1}{1/2} = 2\,. \qquad \square$$

When discussing the convergence of a sequence, we often find it inconvenient to deal with the definition of convergence as given. For this definition makes reference to the number to which the sequence is supposed to converge, and we often do not know this number in advance. Would it not be useful to be able to decide whether a series converges *without knowing to what limit it converges*?

Definition 2.9 Let $\{a_j\}$ be a sequence of real (resp. complex) numbers. We say that the sequence satisfies the *Cauchy criterion* (A. L. Cauchy, 1789–1857)— more briefly, that the sequence is *Cauchy*—if, for each $\epsilon > 0$, there is an integer $N > 0$ such that if $j, k > N$ then $|a_j - a_k| < \epsilon$.

As you study this definition, you will see that it mandates that the elements of the sequence *get close together and stay close together*.

EXAMPLE 2.10 Let $a_j = 1/j$. Of course we know intuitively that this sequence converges to 0. But let us, just for practice, verify that the sequence is Cauchy.

Let $\epsilon > 0$. By the Archimedean principle, choose a positive integer $N > 1/\epsilon$. Then, for $j > k > N$, we have

$$|a_j - a_k| = |1/j - 1/k| = \frac{|j - k|}{jk} < \frac{j}{jk} = \frac{1}{k} < \frac{1}{N} < \epsilon.$$

This shows that the sequence $\{a_j\}$ satisfies the Cauchy criterion. $\qquad\square$

Notice that the concept of a sequence being Cauchy simply makes precise the notion of the elements of the sequence **(i)** *getting* closer together and **(ii)** *staying* close together.

Lemma 2.11 *Every Cauchy sequence is bounded.*

Proof: Let $\epsilon = 1 > 0$. There is an integer $N > 0$ such that $|a_j - a_k| < \epsilon = 1$ whenever $j, k > N$. Thus, if $j \geq N + 1$, we have

$$
\begin{aligned}
|a_j| &\leq |a_{N+1} + (a_j - a_{N+1})| \\
&\leq |a_{N+1}| + |a_j - a_{N+1}| \\
&\leq |a_{N+1}| + 1 \equiv K.
\end{aligned}
$$

Let $L = \max\{|a_1|, |a_2|, \ldots, |a_N|\}$. If j is any natural number, then either $1 \leq j \leq N$, in which case $|a_j| \leq L$, or else $j > N$, in which case $|a_j| \leq K$. Set M = max{K,L}. Then, for any j, $|a_j| \leq M$ as required. $\qquad\square$

In what follows we shall use an interesting and not entirely obvious version of the triangle inequality. You know the triangle inequality as

$$|a + b| \leq |a| + |b|.$$

But let us instead write

$$|a| = |(a + b) - b| = |(a + b) + (-b)| \leq |a + b| + |-b| = |a + b| + |b|.$$

From this we conclude that

$$|a + b| \geq |a| - |b|.$$

A similar argument allows us to analyze $|a - b| < c$. This means that

$$-c < a - b < c$$

or

$$b - c < a$$

and

$$a < b + c.$$

Theorem 2.12 *Let $\{a_j\}$ be a sequence of real numbers. The sequence is Cauchy if and only if it converges to some limit α.*

Proof: First assume that the sequence converges to a limit α. Let $\epsilon > 0$. Choose, by definition of convergence, an integer $N > 0$ such that if $j > N$ then $|a_j - \alpha| < \epsilon/2$. If $j, k > N$ then

$$|a_j - a_k| \leq |a_j - \alpha| + |\alpha - a_k| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.$$

So the sequence is Cauchy.

Conversely, suppose that the sequence is Cauchy. Define

$$S = \{x \in \mathbb{R} : x < a_j \ \text{ for all but finitely many } j\}.$$

[**Hint:** You might find it helpful to think of this set as

$$S = \{x \in \mathbb{R} : \text{there is a positive integer } k \text{ such that } x < a_j \text{ for all } j \geq k\}.]$$

By the lemma, the sequence $\{a_j\}$ is bounded by some positive number M. If x is a real number less than $-M$, then $x \in S$, so S is nonempty. Also S is bounded above by M. Let $\alpha = \sup S$. Then α is a well-defined real number, and we claim that α is the limit of the sequence $\{a_j\}$.

To see this, let $\epsilon > 0$. Choose an integer $N > 0$ such that $|a_j - a_k| < \epsilon/2$ whenever $j, k > N$. Notice that this last inequality implies that

$$|a_j - a_{N+1}| < \epsilon/2 \ \text{ when } \ j \geq N+1 \tag{2.12.1}$$

hence (by the discussion preceding the statement of the theorem)

$$a_j > a_{N+1} - \epsilon/2 \ \text{ when } \ j \geq N+1.$$

Thus $a_{N+1} - \epsilon/2 \in S$ and it follows that

$$\alpha \geq a_{N+1} - \epsilon/2. \tag{2.12.2}$$

Line (2.12.1) also shows that

$$a_j < a_{N+1} + \epsilon/2 \ \text{ when } \ j \geq N+1.$$

Thus $a_{N+1} + \epsilon/2 \notin S$ and

$$\alpha \leq a_{N+1} + \epsilon/2. \tag{2.12.3}$$

Combining lines (2.12.2) and (2.12.3) gives

$$|\alpha - a_{N+1}| \leq \epsilon/2. \tag{2.12.4}$$

But then line (2.12.4) yields, for $j > N$, that

$$|\alpha - a_j| \leq |\alpha - a_{N+1}| + |a_{N+1} - a_j| < \epsilon/2 + \epsilon/2 = \epsilon.$$

This proves that the sequence $\{a_j\}$ converges to α, as claimed. □

Corollary 2.13 *Let $\{\alpha_j\}$ be a sequence of complex numbers. The sequence is Cauchy if and only if it is convergent.*

Proof: Write $\alpha_j = a_j + ib_j$, with a_j, b_j real. Then $\{\alpha_j\}$ is Cauchy if and only if $\{a_j\}$ and $\{b_j\}$ are Cauchy. Also $\{\alpha_j\}$ is convergent to a complex limit α if and only if $\{a_j\}$ converges to $\mathrm{Re}\,\alpha$ and $\{b_j\}$ converges to $\mathrm{Im}\,\alpha$. These observations, together with the theorem, prove the corollary. □

Definition 2.14 *Let $\{a_j\}$ be a sequence of real numbers. The sequence is said to be* increasing *if $a_1 \le a_2 \le \dots$. It is* decreasing *if $a_1 \ge a_2 \ge \dots$.*

A sequence is said to be *monotone* if it is either increasing or decreasing.

EXAMPLE 2.15 Let $a_j = j/(j+1)$. We see that

$$a_j < a_{j+1}$$

just because this is the same as

$$\frac{j}{j+1} < \frac{j+1}{j+2}$$

or

$$j(j+2) < (j+1)^2$$

or

$$j^2 + 2j < j^2 + 2j + 1$$

and that is definitely true. Hence the sequence $\{a_j\}$ is increasing.

On the other hand, let $b_j = (j+1)/j$. We see that

$$b_j > b_{j+1}$$

just because this is the same as

$$\frac{j+1}{j} > \frac{j+2}{j+1}$$

or

$$(j+1)^2 > j(j+2)$$

or

$$j^2 + 2j + 1 > j^2 + 2j$$

and that is definitely true. Hence the sequence $\{b_j\}$ is decreasing. □

Proposition 2.16 *If $\{a_j\}$ is an increasing sequence which is bounded above—$a_j \le M < \infty$ for all j—then $\{a_j\}$ is convergent. If $\{b_j\}$ is a decreasing sequence which is bounded below—$b_j \ge K > -\infty$ for all j—then $\{b_j\}$ is convergent.*

Proof: Let $\epsilon > 0$. Let $\alpha = \sup a_j < \infty$. By definition of supremum, there is an integer N so that $|a_N - \alpha| < \epsilon$. Then, if $\ell \geq N + 1$, we have $a_N \leq a_\ell \leq \alpha$ hence $|a_\ell - \alpha| < \epsilon$. Thus the sequence converges to α.

The proof for decreasing sequences is similar and we omit it. □

EXAMPLE 2.17 Let $a_1 = \sqrt{2}$ and set $a_{j+1} = \sqrt{2 + a_j}$ for $j \geq 1$. You can verify that $\{a_j\}$ is increasing and bounded above (by 4 for example). What is its limit (which is guaranteed to exist by the proposition)? □

A proof very similar to that of the proposition gives the following useful fact:

Corollary 2.18 *Let S be a nonempty set of real numbers which is bounded above and below. Let β be its supremum and α its infimum. If $\epsilon > 0$ then there are $s, t \in S$ such that $|s - \beta| < \epsilon$ and $|t - \alpha| < \epsilon$.*

Proof: This is a restatement of the proof of the proposition. □

EXAMPLE 2.19 Let S be the set $(0, 2) \subseteq \mathbb{R}$. Then the supremum of S is 2. And, if $\epsilon > 0$ is small, then the point $\alpha_\epsilon = 2 - \epsilon/2$ lies in the set. Note that $|\alpha_\epsilon - 2| < \epsilon$.

Likewise, the infimum of S is 0. And, if $\epsilon > 0$ is small, then the point $\beta_\epsilon = 0 + \epsilon/2$ lies in the set. Note that $|\beta_\epsilon - 0| < \epsilon$. □

We conclude the section by recording one of the most useful results for calculating the limit of a sequence:

Proposition 2.20 (The Pinching Principle) *Let $\{a_j\}, \{b_j\}$, and $\{c_j\}$ be sequences of real numbers satisfying*

$$a_j \leq b_j \leq c_j$$

for every j sufficiently large. If

$$\lim_{j \to \infty} a_j = \lim_{j \to \infty} c_j = \alpha$$

for some real number α, then

$$\lim_{j \to \infty} b_j = \alpha \,.$$

Proof: This proof is requested of you in the exercises. □

EXAMPLE 2.21 Define
$$a_j = \frac{\sin j \cos 2j}{j^2}.$$

Then
$$0 \le |a_j| \le \frac{1}{j^2}.$$

It is clear that
$$\lim_{j \to \infty} 0 = 0$$

and
$$\lim_{j \to \infty} \frac{1}{j^2} = 0.$$

Therefore
$$\lim_{j \to \infty} |a_j| = 0$$

so that
$$\lim_{j \to \infty} a_j = 0. \qquad \square$$

Exercises

1. Suppose a sequence $\{a_j\}$ has the property that, for every natural number N, there is a j_N such that $a_{j_N} = a_{j_{N+1}} = \cdots = a_{j_N+N}$. In other words, the sequence has arbitrarily long repetitive strings. Does it follow that the sequence converges?

2. Let α be an irrational real number and let a_j be a sequence of rational numbers converging to α. Suppose that each a_j is a fraction expressed in lowest terms: $a_j = \alpha_j/\beta_j$. Prove that the β_j are unbounded.

3. Let $\{a_j\}$ be a sequence of rational numbers all of which have denominator a power of 2. What are the possible limits of such a sequence?

4. Redo Exercise **3** with the additional hypothesis that all of the denominators are less than or equal to 2^{10}.

5. Use the integral of $1/(1 + t^2)$, together with Riemann sums (ideas which you know from calculus, and which we shall treat rigorously later in the book), to develop a scheme for calculating the digits of π.

6. Prove Corollary 2.18.

7. Prove Proposition 2.20.

8. Prove parts **(2)** and **(4)** of Proposition 2.6.

9. Give an example of a decreasing sequence that converges to π.

10. Prove the following result, which we have used without comment in the text: Let S be a set of real numbers which is bounded above and let $t = \sup S$. For any $\epsilon > 0$ there is an element $s \in S$ such that $t - \epsilon < s \leq t$. (**Remark:** Notice that this result makes good intuitive sense: the elements of S should become arbitrarily close to the supremum t, otherwise there would be enough room to decrease the value of t and make the supremum even smaller.) Formulate and prove a similar result for the infimum.

11. Let $\{a_j\}$ be a sequence of real or complex numbers. Suppose that every subsequence has itself a subsequence which converges to a given number α. Prove that the full sequence converges to α.

* **12.** Let $\{a_j\}$ be a sequence of complex numbers. Suppose that, for every pair of integers $N > M > 0$, it holds that $|a_M - a_{M+1}| + |a_{M+1} - a_{M+2}| + \cdots + |a_{N-1} - a_N| \leq 1$. Prove that $\{a_j\}$ converges.

13. Let $a_1, a_2 > 0$ and for $j \geq 3$ define $a_j = a_{j-1} + a_{j-2}$. Show that this sequence cannot converge to a finite limit.

2.2   Subsequences

Let $\{a_j\}$ be a given sequence. If

$$0 < j_1 < j_2 < \cdots$$

are positive integers then the function

$$k \mapsto a_{j_k}$$

is called a *subsequence* of the given sequence. We usually write the subsequence as

$$\{a_{j_k}\}_{k=1}^{\infty} \quad \text{or} \quad \{a_{j_k}\}.$$

EXAMPLE 2.22 Consider the sequence

$$\{2^j\} = \{2, 4, 8, \ldots\}.$$

Then the sequence

$$\{2^{2k}\} = \{4, 16, 64, \ldots\} \tag{2.22.1}$$

is a subsequence. Notice that the subsequence contains a subcollection of elements of the original sequence *in the same order*. In this example, $j_k = 2k$.

Another subsequence is

$$\{2^{(2^k)}\} = \{4, 16, 256, \ldots\}. \tag{2.22.2}$$

In this instance, it holds that $j_k = 2^k$. Notice that this new subsequence is in fact a subsequence of the first subsequence (2.22.1). That is, it is a sub-subsequence of the original sequence $\{2^j\}$.   □

Proposition 2.23 *If $\{a_j\}$ is a convergent sequence with limit α, then every subsequence converges to the limit α.*

Conversely, if a sequence $\{b_j\}$ has the property that each of its subsequences is convergent then $\{b_j\}$ itself is convergent.

Proof: Assume $\{a_j\}$ is convergent to a limit α, and let $\{a_{j_k}\}$ be a subsequence. Let $\epsilon > 0$ and choose $N > 0$ such that $|a_j - \alpha| < \epsilon$ whenever $j > N$. Now if $k > N$ then $j_k > N$ hence $|a_{j_k} - \alpha| < \epsilon$. Therefore, by definition, the subsequence $\{a_{j_k}\}$ also converges to α.

The converse is trivial, simply because the sequence is a subsequence of itself. □

Now we present one of the most fundamental theorems of basic real analysis (due to B. Bolzano, 1781–1848, and K. Weierstrass, 1815–1897).

Theorem 2.24 (Bolzano–Weierstrass) *Let $\{a_j\}$ be a bounded sequence in $\mathbb{R}$. Then there is a subsequence which converges.*

Proof: Suppose that $|a_j| \leq M$ for every j. We may assume that $M > 0$. It is convenient to formulate our hypothesis as $a_j \in [-M, M]$ for every j.

One of the two intervals $[-M, 0]$ and $[0, M]$ must contain infinitely many elements of the sequence. Assume that $[0, M]$ does. Choose a_{j_1} to be one of the infinitely many sequence elements in $[0, M]$.

Next, one of the intervals $[0, M/2]$ and $[M/2, M]$ must contain infinitely many elements of the sequence. Suppose that it is $[0, M/2]$. Choose an element a_{j_2}, with $j_2 > j_1$, from $[0, M/2]$. Continue in this fashion, halving the interval, choosing a half with infinitely many sequence elements, and selecting the next subsequential element from that half.

Let us analyze the resulting subsequence. Notice that $|a_{j_1} - a_{j_2}| \leq M$ since both elements belong to the interval $[0, M]$. Likewise, $|a_{j_2} - a_{j_3}| \leq M/2$ since both elements belong to $[0, M/2]$. In general, $|a_{j_k} - a_{j_{k+1}}| \leq 2^{-k+1} \cdot M$ for each $k \in \mathbb{N}$.

Now let $\epsilon > 0$. Choose an integer $N > 0$ such that $2^{-N} < \epsilon/(4M)$. Then, for any $m > l > N$ we have

$$
\begin{aligned}
|a_{j_l} - a_{j_m}| &= |(a_{j_l} - a_{j_{l+1}}) + (a_{j_{l+1}} - a_{j_{l+2}}) + \cdots + (a_{j_{m-1}} - a_{j_m})| \\
&\leq |a_{j_l} - a_{j_{l+1}}| + |a_{j_{l+1}} - a_{j_{l+2}}| + \cdots + |a_{j_{m-1}} - a_{j_m}| \\
&\leq 2^{-l+1} \cdot M + 2^{-l} \cdot M + \cdots + 2^{-m+2} \cdot M \\
&= \left(2^{-l+1} + 2^{-l} + \cdots + 2^{-m+2}\right) \cdot M \\
&= \left((2^{-l+2} - 2^{-l+1}) + (2^{-l+1} - 2^{-l}) + \cdots \right. \\
&\qquad \left. + (2^{-m+3} - 2^{-m+2})\right) \cdot M \\
&= \left(2^{-l+2} - 2^{-m+2}\right) \cdot M \\
&< 2^{-l+2} \cdot M \\
&< 4 \cdot \frac{\epsilon}{4M} \cdot M \\
&= \epsilon.
\end{aligned}
$$

We see that the subsequence $\{a_{j_k}\}$ is Cauchy, so it converges. □

Remark 2.25 The Bolzano–Weierstrass theorem is a generalization of our result from the last section about increasing sequences which are bounded above (resp. decreasing sequences which are bounded below). For such a sequence is surely bounded above *and* below (why?). So it has a convergent subsequence. And thus it follows easily that the entire sequence converges. Details are left as an exercise.

It is a fact—which you can verify for yourself—that *any* real sequence has a monotone subsequence. This observation implies Bolzano–Weierstrass.

EXAMPLE 2.26 In this text we have not yet given a rigorous definition of the function $\sin x$ (see Section 9.3). However, just for the moment, use the definition you learned in calculus class and consider the sequence $\{\sin j\}_{j=1}^{\infty}$. Notice that the sequence is bounded in absolute value by 1. The Bolzano–Weierstrass theorem guarantees that there is a convergent subsequence, even though it would be very difficult to say precisely what that convergent subsequence is. □

Corollary 2.27 *Let $\{\alpha_j\}$ be a bounded sequence of complex numbers. Then there is a convergent subsequence.*

Proof: Write $\alpha_j = a_j + ib_j$, with $a_j, b_j \in \mathbb{R}$. The fact that $\{\alpha_j\}$ is bounded implies that $\{a_j\}$ is bounded. By the Bolzano–Weierstrass theorem, there is a convergent subsequence $\{a_{j_k}\}$.

Now the sequence $\{b_{j_k}\}$ is bounded. So it has a convergent subsequence $\{b_{j_{k_l}}\}$. Then the sequence $\{\alpha_{j_{k_l}}\}$ is convergent, and is a subsequence of the original sequence $\{\alpha_j\}$. □

In earlier parts of this chapter we have discussed sequences that converge to a finite number. Such a sequence is, by Proposition 2.5, bounded. However, in some mathematical contexts, it is useful to speak of a sequence "diverging[1] to infinity." We now will treat briefly the idea of "divergence to infinity."

Definition 2.28 We say that a sequence $\{a_j\}$ of real numbers *diverges to* $+\infty$ if, for every $M > 0$, there is an integer $N > 0$ such that $a_j > M$ whenever $j > N$. We write $a_j \to +\infty$.

We say that $\{a_j\}$ *diverges to* $-\infty$ if, for every $K > 0$, there is an integer $N > 0$ such that $a_j < -K$ whenever $j > N$. We write $a_j \to -\infty$.

Remark 2.29 Notice that the statement $a_j \to +\infty$ means that we can make a_j become arbitrarily large and positive and *stay* large and positive just by making j large enough.

Likewise, the statement $a_j \to -\infty$ means that we can force a_j to be arbitrarily large and negative, and *stay* large and negative, just by making j large enough.

EXAMPLE 2.30 The sequence $\{j^2\}$ diverges to $+\infty$. The sequence $\{-2j + 18\}$ diverges to $-\infty$. The sequence $\{j + (-1)^j \cdot j\}$ has no infinite limit and no finite limit. However, the subsequence $\{0, 0, 0, \dots\}$ converges to 0 and the subsequence $\{4, 8, 12 \dots\}$ diverges to $+\infty$. □

With the new language provided by Definition 2.28, we may generalize Proposition 2.16:

Proposition 2.31 *Let* $\{a_j\}$ *be an increasing sequence of real numbers. Then the sequence has a limit—either a finite number or* $+\infty$.

Let $\{b_j\}$ *be a decreasing sequence of real numbers. Then the sequence has a limit—either a finite number or* $-\infty$.

In the same spirit as the last definition, we also have the following:

Definition 2.32 If S is a set of real numbers which is *not* bounded above, we say that its supremum (or least upper bound) is $+\infty$.

If T is a set of real numbers which is *not* bounded below, then we say that its infimum (or greatest lower bound) is $-\infty$.

Exercises

1. Use the Bolzano–Weierstrass theorem to show that every decreasing sequence that is bounded below converges.

---

[1]Some books say "converging to infinity," but this terminology can be confusing.

2. Give an example of a sequence of rational numbers with the property that, for any real number α, or for $\alpha = +\infty$ or $\alpha = -\infty$, there is a subsequence approaching α.

3. Prove that if $\{a_j\}$ has a subsequence diverging to $\pm\infty$ then $\{a_j\}$ cannot converge.

4. Let $x_1 = 2$. For $j \geq 1$, set

$$x_{j+1} = x_j - \frac{x_j^2 - 2}{2x_j}.$$

Show that the sequence $\{x_j\}$ is decreasing and bounded below. What is its limit?

5. The sequence

$$a_j = (1 + 1/2 + 1/3 + \cdots + 1/j) - \log j$$

is a famous example. It is known to converge, but nobody knows whether the limit is rational or irrational. Draw a picture which shows that the sequence converges.

6. Provide the details of the proof of Proposition 2.31.

* 7. Provide the details of the assertion that the sequence $\{\cos j\}$ is dense in the interval $[-1, 1]$.

* 8. Let n be a positive integer. Consider n, $n + 1$, ... modulo π. This means that you subtract from each number the greatest multiple of π that does not exceed it. Prove that this collection of numbers is dense in $[0, \pi]$. That is, the numbers get arbitrarly close to any element of this interval.

* 9. Let $S = \{0, 1, 1/2, 1/3, 1/4, \ldots\}$. Give an example of a sequence $\{a_j\}$ with the property that, for each $s \in S$, there is a subsequence converging to s, but no subsequence converges to any limit not in S.

* 10. Give another proof of the Bolzano–Weierstrass theorem as follows. If $\{a_j\}$ is a bounded sequence let $b_j = \inf\{a_j, a_{j+1}, \ldots\}$. Then each b_j is finite, $b_1 \leq b_2 \leq \ldots$, and $\{b_j\}$ is bounded above. Now use Proposition 2.16.

* 11. Prove that the sequence

$$a_N = \sum_{m=1}^{N} \frac{\sin m}{m}$$

converges.

* 12. Prove that the sequence

$$a_N = \sum_{m=1}^{N} \frac{\sin^2 m}{m}$$

diverges.

2.3   Lim sup and Lim inf

Convergent sequences are useful objects, but the unfortunate truth is that most sequences do not converge. Nevertheless, we would like to have a language for discussing the asymptotic behavior of *any* real sequence $\{a_j\}$ as $j \to \infty$. That is the purpose of the concepts of "limit superior" (or "upper limit") and "limit inferior" (or "lower limit").

Definition 2.33 Let $\{a_j\}$ be a sequence of real numbers. For each j let

$$A_j = \inf\{a_j, a_{j+1}, a_{j+2}, \dots\}.$$

Then $\{A_j\}$ is an increasing sequence (since, as j becomes large, we are taking the infimum of a smaller set of numbers), so it has a limit (either a finite limit or $\pm\infty$). We define the *limit infimum* of $\{a_j\}$ to be

$$\liminf a_j = \lim_{j \to \infty} A_j\,.$$

It is common to refer to this number as the lim inf of the sequence.

Likewise, let

$$B_j = \sup\{a_j, a_{j+1}, a_{j+2}, \dots\}\,.$$

Then $\{B_j\}$ is a decreasing sequence (since, as j becomes large, we are taking the supremum of a smaller set of numbers), so it has a limit (either a finite limit of $\pm\infty$). We define the *limit supremum* of $\{a_j\}$ to be

$$\limsup a_j = \lim_{j \to \infty} B_j.$$

It is common to refer to this number as the lim sup of the sequence.

Notice that the lim sup or lim inf of a sequence can be $\pm\infty$.

Remark 2.34 What is the intuitive content of this definition? For each j, A_j picks out the greatest lower bound of the sequence in the jth position or later. So the sequence $\{A_j\}$ should tend to the *smallest* possible limit of any subsequence of $\{a_j\}$.

Likewise, for each j, B_j picks out the least upper bound of the sequence in the jth position or later. So the sequence $\{B_j\}$ should tend to the *greatest* possible limit of any subsequence of $\{a_j\}$. We shall make these remarks more precise in Proposition 2.36 below.

Notice that it is implicit in the definition that *every* real sequence has a limit supremum and a limit infimum.

EXAMPLE 2.35 Consider the sequence $\{(-1)^j\}$. Of course this sequence does not converge. Let us calculate its lim sup and lim inf.

Referring to the definition, we have that $A_j = -1$ for every j. So

$$\liminf (-1)^j = \lim (-1) = -1.$$

Similarly, $B_j = +1$ for every j. Therefore

$$\limsup{(-1)^j} = \lim{(+1)} = +1.$$

As we predicted in the remark, the $\liminf$ is the least subsequential limit, and the $\limsup$ is the greatest subsequential limit. $\square$

Now let us prove the characterizing property of $\limsup$ and $\liminf$ to which we have been alluding.

Proposition 2.36 *Let $\{a_j\}$ be a sequence of real numbers. Let $\beta = \limsup_{j \to \infty} a_j$ and $\alpha = \liminf_{j \to \infty} a_j$. If $\{a_{j_\ell}\}$ is any subsequence of the given sequence then*

$$\alpha \leq \liminf_{\ell \to \infty} a_{j_\ell} \leq \limsup_{\ell \to \infty} a_{j_\ell} \leq \beta.$$

Moreover, there is a subsequence $\{a_{j_k}\}$ such that

$$\lim_{k \to \infty} a_{j_k} = \alpha$$

and another sequence $\{a_{j_m}\}$ such that

$$\lim_{k \to \infty} a_{j_m} = \beta.$$

Proof: For simplicity in this proof we assume that the $\limsup$ and $\liminf$ are finite. The case of infinite $\limsup$s and $\liminf$s is treated in the exercises.

We begin by considering the $\liminf$. There is a $j_1 \geq 1$ such that $|A_1 - a_{j_1}| < 2^{-1}$. We choose j_1 to be as small as possible. Next, we choose j_2, necessarily greater than j_1, such that j_2 is as small as possible and $|a_{j_2} - A_2| < 2^{-2}$. Continuing in this fashion, we select $j_k > j_{k-1}$ such that $|a_{j_k} - A_k| < 2^{-k}$, etc.

Recall that $A_k \to \alpha = \liminf_{j \to \infty} a_j$. Now fix $\epsilon > 0$. If N is an integer so large that $k > N$ implies that $|A_k - \alpha| < \epsilon/2$ and also that $2^{-N} < \epsilon/2$ then, for such k, we have

$$
\begin{aligned}
|a_{j_k} - \alpha| &\leq & |a_{j_k} - A_k| + |A_k - \alpha| \\
&<& 2^{-k} + \frac{\epsilon}{2} \\
&<& \frac{\epsilon}{2} + \frac{\epsilon}{2} \\
&=& \epsilon.
\end{aligned}
$$

Thus the subsequence $\{a_{j_k}\}$ converges to α, the $\liminf$ of the given sequence. A similar construction gives a (different) subsequence $\{a_{n_k}\}$ converging to β, the $\limsup$ of the given sequence.

Now let $\{a_{j_\ell}\}$ be *any* subsequence of the sequence $\{a_j\}$. Let β^* be the $\limsup$ of this subsequence. Then, by the first part of the proof, there is a subsequence $\{a_{j_{\ell_m}}\}$ such that

$$\lim_{m \to \infty} a_{j_{\ell_m}} = \beta^*.$$

But $a_{j_{\ell_m}} \le B_{j_{\ell_m}}$ by the very definition of the Bs. Thus

$$\beta^* = \lim_{m \to \infty} a_{j_{\ell_m}} \le \lim_{m \to \infty} B_{j_{\ell_m}} = \beta$$

or

$$\limsup_{\ell \to \infty} a_{j_\ell} \le \beta \,,$$

as claimed. A similar argument shows that

$$\liminf_{l \to \infty} a_{j_l} \ge \alpha \,.$$

This completes the proof of the proposition. $\qquad\qquad\square$

Corollary 2.37 *If $\{a_j\}$ is a sequence and $\{a_{j_k}\}$ is a convergent subsequence then*

$$\liminf_{j \to \infty} a_j \le \lim_{k \to \infty} a_{j_k} \le \limsup_{j \to \infty} a_j \,.$$

EXAMPLE 2.38 Consider the sequence

$$a_j = \frac{(-1)^j j^2}{j^2 + j} \,.$$

It is helpful to rewrite this sequence as

$$a_j = (-1)^j \cdot \left(1 - \frac{j}{j^2 + j}\right) \,.$$

Then, looking at the terms of even index, it is easy to see that the lim sup of this sequence is $+1$. And, looking at the terms of odd index, it is easy to see that the lim inf of this sequence is -1.

Every convergent subsequence of $\{a_j\}$ will have limit lying between -1 and $+1$. $\qquad\qquad\square$

We close this section with a fact that is analogous to one for the supremum and infimum. Its proof is analogous to arguments we have seen before.

Proposition 2.39 *Let $\{a_j\}$ be a sequence and set $\limsup a_j = \beta$ and $\liminf a_j = \alpha$. Assume that α, β are finite real numbers. Let $\epsilon > 0$. Then there are arbitrarily large j such that $a_j > \beta - \epsilon$. Also there are arbitrarily large k such that $a_k < \alpha + \epsilon$.*

EXAMPLE 2.40 Consider the sequence $\{a_j\}$ in the last example. Let $\epsilon > 0$. Choose j even so that $j > (1 - \epsilon)/\epsilon$. Then

$$\frac{j^2}{j^2 + j} > 1 - \epsilon \,.$$

Now again choose $\epsilon > 0$. Choose j odd so that $j > (1 - \epsilon)/\epsilon$. Then

$$-\frac{j^2}{j^2 + j} < -1 + \epsilon \,. \qquad\qquad\square$$

Exercises

1. Consider $\{a_j\}$ both as a sequence and as a set. How are the lim sup and the sup related? How are the lim inf and the inf related? Give examples.

2. Let $\{a_j\}$ be a sequence of positive numbers. How are the lim sup and lim inf of $\{a_j\}$ related to the lim sup and lim inf of $\{1/a_j\}$?

3. How are the lim sup and lim inf of $\{a_j\}$ related to the lim sup and lim inf of $\{-a_j\}$?

4. Let $\{a_j\}$ be a real sequence. Prove that if

$$\liminf a_j = \limsup a_j$$

   then the sequence $\{a_j\}$ converges. Prove the converse as well.

5. Let $a < b$ be real numbers. Give an example of a real sequence whose lim sup is b and whose lim inf is a.

6. Explain why we can make no sense of the concepts of lim sup and lim inf for complex sequences.

7. Let $\{a_j\}, \{b_j\}$ be sequences of real numbers. Prove the inequality $\limsup(a_j + b_j) \leq \limsup a_j + \limsup b_j$. How are the lim infs related? How is the quantity $(\limsup a_j) \cdot (\limsup b_j)$ related to $\limsup(a_j \cdot b_j)$? How are the lim infs related?

8. Give an example of a sequence whose lim sup and lim inf differ by 1.

9. Prove Corollary 2.37.

10. Prove Proposition 2.39.

11. Prove a version of Proposition 2.36 when the indicated lim sup and/or lim inf are $\pm\infty$.

12. Prove a version of Proposition 2.39 when the indicated lim sup and/or lim inf are $\pm\infty$.

* 13. Find the lim sup and lim inf of the sequences

$$\{|\sin j|^{\sin j}\} \quad \text{and} \quad \{|\cos j|^{\cos j}\}.$$

2.4   Some Special Sequences

We often obtain information about a new sequence by comparison with a sequence that we already know. Thus it is well to have a catalogue of fundamental sequences which provide a basis for comparison.

EXAMPLE 2.41 Fix a real number a. The sequence $\{a^j\}$ is called a *power se-quence*. If $-1 < a < 1$ then the sequence converges to 0. If $a = 1$ then the sequence is a constant sequence and converges to 1. If $a > 1$ then the sequence diverges to $+\infty$. Finally, if $a \leq -1$ then the sequence diverges. $\square$

Recall that, in Section 1.1, we discussed the existence of nth roots of positive real numbers. If $\alpha > 0, m \in \mathbb{Z}$, and $n \in \mathbb{N}$ then we may define

$$\alpha^{m/n} = (\alpha^m)^{1/n}.$$

Thus we may talk about rational powers of a positive number. Next, if $\beta \in \mathbb{R}$ then we may define

$$\alpha^\beta = \sup\{\alpha^q : q \in \mathbb{Q}, q < \beta\}.$$

Thus we can define *any real power* of a positive real number. The exercises ask you to verify several basic properties of these exponentials.

Lemma 2.42 *If $\alpha > 1$ is a real number and $\beta > 0$ then $\alpha^\beta > 1$.*

Proof: Let q be a positive rational number which is less than β. Suppose that $q = m/n$, with m, n integers. It is obvious that $\alpha^m > 1$ and hence that $(\alpha^m)^{1/n} > 1$. Since α^β majorizes this last quantity, we are done. $\square$

EXAMPLE 2.43 Fix a real number α and consider the sequence $\{j^\alpha\}$. If $\alpha > 0$ then it is easy to see that $j^\alpha \to +\infty$: to verify this assertion fix $M > 0$ and take the number N to be the first integer after $M^{1/\alpha}$.

If $\alpha = 0$ then j^α is a constant sequence, identically equal to 1.

If $\alpha < 0$ then $j^\alpha = 1/j^{-\alpha}$. The denominator of this last expression tends to $+\infty$ hence the sequence j^α tends to 0. $\square$

EXAMPLE 2.44 The sequence $\{j^{1/j}\}$ converges to 1. In fact, consider the expressions $\alpha_j = j^{1/j} - 1 > 0$. We have that

$$j = (\alpha_j + 1)^j \geq \frac{j(j-1)}{2}(\alpha_j)^2,$$

(the latter being just one term from the binomial expansion). Thus

$$0 < \alpha_j \leq \sqrt{2/(j-1)}$$

as long as $j \geq 2$. It follows that $\alpha_j \to 0$ or $j^{1/j} \to 1$. $\square$

EXAMPLE 2.45 Let α be a positive real number. Then the sequence $\alpha^{1/j}$ converges to 1. To see this, first note that the case $\alpha = 1$ is trivial, and the case $\alpha > 1$ implies the case $\alpha < 1$ (by taking reciprocals). So we concentrate on $\alpha > 1$. But then we have

$$1 < \alpha^{1/j} < j^{1/j}$$

when $j > \alpha$. Since $j^{1/j}$ tends to 1, Proposition 2.20 applies and the proof is complete. $\square$

EXAMPLE 2.46 Let $\lambda > 1$ and let α be real. Then the sequence

$$\left\{ \frac{j^\alpha}{\lambda^j} \right\}_{j=1}^\infty$$

converges to 0.

To see this, fix an integer $k > \alpha$ and consider $j > 2k$. [Notice that k is fixed once and for all but j will be allowed to tend to $+\infty$ at the appropriate moment.] Writing $\lambda = 1 + \mu, \mu > 0$, we have that

$$\lambda^j = (1 + \mu)^j > \frac{j(j-1)(j-2)\cdots(j-k+1)}{k(k-1)(k-2)\cdots 2 \cdot 1} \cdot 1^{j-k} \cdot \mu^k \,.$$

Of course this comes from picking out the kth term of the binomial expansion for $(1 + \mu)^j$. Notice that, since $j > 2k$, then each of the expressions $j, (j-1), \ldots (j-k+1)$ in the numerator on the right exceeds $j/2$. Thus

$$\lambda^j > \frac{j^k}{2^k \cdot k!} \cdot \mu^k$$

and

$$0 < \frac{j^\alpha}{\lambda^j} < j^\alpha \cdot \frac{2^k \cdot k!}{j^k \cdot \mu^k} = \frac{j^{\alpha-k} \cdot 2^k \cdot k!}{\mu^k} \,.$$

Since $\alpha - k < 0$, the right side tends to 0 as $j \to \infty$.   □

EXAMPLE 2.47 The sequence

$$\left\{ \left(1 + \frac{1}{j} \right)^j \right\}$$

converges. In fact it is increasing and bounded above. Use the Binomial Expansion to prove this assertion. The limit of the sequence is the number that we shall later call e (in honor of Leonhard Euler, 1707–1783, who first studied it in detail). We shall study this sequence in detail later in the book.   □

EXAMPLE 2.48 The sequence

$$\left(1 - \frac{1}{j} \right)^j$$

converges to $1/e$, where the definition of e is given in the last example. More generally, the sequence

$$\left(1 + \frac{x}{j} \right)^j$$

converges to e^x (here e^x is defined as in the discussion following Example 2.41 above).   □

Exercises

1. Let α be a positive real number and let $p/q = m/n$ be two different representations of the same rational number r. Prove that
$$(\alpha^m)^{1/n} = (\alpha^p)^{1/q}.$$
   Also prove that
$$(\alpha^{1/n})^m = (\alpha^m)^{1/n}.$$
   If β is another positive real and γ is any real then prove that
$$(\alpha \cdot \beta)^\gamma = \alpha^\gamma \cdot \beta^\gamma.$$

2. Discuss the convergence of the sequence $\{(1/j)^{1/j}\}_{j=1}^{\infty}$.

3. Discuss the convergence of the sequence $\{(j^j)/(2j)!\}_{j=2}^{\infty}$.

4. Prove that the exponential, as defined in this section, satisfies
$$(a^b)^c = a^{bc} \qquad \text{and} \qquad a^b a^c = a^{b+c}.$$

* **5.** Refer to Exercise 5 in Section 2.2. Consider the sequence given by
$$a_j = \left[1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{j}\right] - \log j.$$
   Then $\{a_j\}$ converges to a limit γ. This number was first studied by Euler. It arises in many different contexts in analysis and number theory.

   As a challenge problem, show that
$$|a_j - \gamma| \leq \frac{C}{j}$$
   for some universal constant $C > 0$.

* **6.** Give a recursive definition of the Fibonacci sequence. Find a generating function for the Fibonacci sequence and use it to derive an explicit formula for the nth term of the sequence.

7. A sequence is defined by the rule $a_0 = 2$, $a_1 = 1$, and $a_j = 3a_{j-1} - a_{j-2}$. Find a formula for a_j.

8. A sequence is defined by the rule $a_0 = 4$, $a_1 = -1$, and $a_j = -a_{j-1} + 2a_{j-2}$. Find a formula for a_j.

* **9.** Consider the sequence
$$a_j = \left(1 + \frac{1}{1^2}\right) \cdot \left(1 + \frac{1}{2^2}\right) \cdot \left(1 + \frac{1}{3^2}\right) \cdot \cdots \cdot \left(1 + \frac{1}{j^2}\right).$$
   Discuss convergence and divergence.

*   **10.** Prove that
$$\left(1 + \frac{x}{j}\right)^{j}$$
converges to e^x for any real number x.

Chapter 3

Series of Numbers

3.1 Convergence of Series

In this section we will use standard summation notation:

$$\sum_{j=m}^{n} a_j \equiv a_m + a_{m+1} + \cdots + a_n \,.$$

A series is an infinite sum. One of the most effective ways to handle an infinite process in mathematics is with a limit. This consideration leads to the following definition:

Definition 3.1 The formal expression

$$\sum_{j=1}^{\infty} a_j \,,$$

where the a_js are real or complex numbers, is called a *series*. For $N = 1, 2, 3, \ldots$, the expression

$$S_N = \sum_{j=1}^{N} a_j = a_1 + a_2 + \ldots a_N$$

is called the Nth *partial sum* of the series. In case

$$\lim_{N \to \infty} S_N$$

exists and is finite we say that the series *converges*. The limit of the partial sums is called the *sum* of the series. If the series does not converge, then we say that the series *diverges*.

Notice that the question of convergence of a series, which should be thought of as an *addition process*, reduces to a question about the *sequence* of partial sums.

EXAMPLE 3.2 Consider the series

$$\sum_{j=1}^{\infty} 2^{-j} \, .$$

The Nth partial sum for this series is

$$S_N = 2^{-1} + 2^{-2} + \cdots + 2^{-N} \, .$$

In order to determine whether the sequence $\{S_N\}$ has a limit, we rewrite S_N as

$$
\begin{aligned}
S_N \;\; &= \;\; \left(2^{-0} - 2^{-1}\right) + \left(2^{-1} - 2^{-2}\right) + \ldots \\
& \qquad \left(2^{-N+1} - 2^{-N}\right) \, .
\end{aligned}
$$

The expression on the right of the last equation telescopes (i.e., successive pairs of terms cancel) and we find that

$$S_N = 2^{-0} - 2^{-N} \, .$$

Thus

$$\lim_{N \to \infty} S_N = 2^{-0} = 1 \, .$$

We conclude that the series converges.   □

EXAMPLE 3.3 Let us examine the series

$$\sum_{j=1}^{\infty} \frac{1}{j}$$

for convergence or divergence. (This series is commonly called the *harmonic series* because it describes the harmonics in music.) Now

$$
\begin{aligned}
S_1 \;\; &= \;\; 1 = \frac{2}{2} \\
S_2 \;\; &= \;\; 1 + \frac{1}{2} = \frac{3}{2} \\
S_4 \;\; &= \;\; 1 + \frac{1}{2} + \left(\frac{1}{3} + \frac{1}{4}\right) \\
&\geq \;\; 1 + \frac{1}{2} + \left(\frac{1}{4} + \frac{1}{4}\right) \geq 1 + \frac{1}{2} + \frac{1}{2} = \frac{4}{2} \\
S_8 \;\; &= \;\; 1 + \frac{1}{2} + \left(\frac{1}{3} + \frac{1}{4}\right) + \left(\frac{1}{5} + \frac{1}{6} + \frac{1}{7} + \frac{1}{8}\right) \\
&\geq \;\; 1 + \frac{1}{2} + \left(\frac{1}{4} + \frac{1}{4}\right) + \left(\frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8}\right) \\
&= \;\; \frac{5}{2} \, .
\end{aligned}
$$

In general this argument shows that

$$S_{2^k} \geq \frac{k+2}{2} \, .$$

The sequence of S_Ns is increasing since the series contains only positive terms. The fact that the partial sums $S_1, S_2, S_4, S_8, \ldots$ increases without bound shows that the entire sequence of partial sums must increase without bound. We conclude that the series diverges. □

Just as with sequences, we have a Cauchy criterion for series:

Proposition 3.4 *The series $\sum_{j=1}^{\infty} a_j$ converges if and only if, for every $\epsilon > 0$, there is an integer $N \geq 1$ such that, if $n \geq m > N$, then*

$$\left| \sum_{j=m}^{n} a_j \right| < \epsilon. \tag{3.4.1}$$

The condition (3.4.1) is called the Cauchy criterion for series.

Proof: Suppose that the Cauchy criterion holds. Pick $\epsilon > 0$ and choose N so large that (3.4.1) holds. If $n \geq m > N$, then

$$|S_n - S_m| = \left| \sum_{j=m+1}^{n} a_j \right| < \epsilon$$

by hypothesis. Thus the sequence $\{S_N\}$ is Cauchy in the sense discussed for sequences in Section 2.1. We conclude that the sequence $\{S_N\}$ converges; by definition, therefore, the series converges.

Conversely, if the series converges then, by definition, the sequence $\{S_N\}$ of partial sums converges. In particular, the sequence $\{S_N\}$ must be Cauchy. Thus, for any $\epsilon > 0$, there is a number $N > 0$ such that if $n \geq m > N$ then

$$|S_n - S_m| < \epsilon \, .$$

This just says that

$$\left| \sum_{j=m+1}^{n} a_j \right| < \epsilon \, ,$$

and this last inequality is the Cauchy criterion for series. □

EXAMPLE 3.5 Let us use the Cauchy criterion to verify that the series

$$\sum_{j=1}^{\infty} \frac{1}{j \cdot (j+1)}$$

converges.

Notice that, if $n \geq m > 1$, then

$$\left| \sum_{j=m}^{n} \frac{1}{j \cdot (j+1)} \right| = \left(\frac{1}{m} - \frac{1}{m+1} \right) + \left(\frac{1}{m+1} - \frac{1}{m+2} \right) + \cdots$$
$$+ \left(\frac{1}{n} - \frac{1}{n+1} \right).$$

The sum on the right plainly telescopes and we have

$$\left| \sum_{j=m}^{n} \frac{1}{j \cdot (j+1)} \right| = \frac{1}{m} - \frac{1}{n+1}.$$

Let $\epsilon > 0$. Let us choose N to be the next integer after $1/\epsilon$. Then, for $n \geq m > N$, we may conclude that

$$\left| \sum_{j=m}^{n} \frac{1}{j \cdot (j+1)} \right| = \frac{1}{m} - \frac{1}{n+1} < \frac{1}{m} < \frac{1}{N} < \epsilon.$$

This is the desired conclusion.   □

The next result gives a necessary condition for a series to converge. It is a useful device for detecting divergent series, although it can never tell us that a series converges.

Proposition 3.6 (The Zero Test) *If the series*

$$\sum_{j=1}^{\infty} a_j$$

converges then the terms a_j tend to zero as $j \to \infty$.

Proof: Since we are assuming that the series converges, then it must satisfy the Cauchy criterion. Let $\epsilon > 0$. Then there is an integer $N \geq 1$ such that, if $n \geq m > N$, then

$$\left| \sum_{j=m}^{n} a_j \right| < \epsilon. \tag{3.6.1}$$

We take $n = m$ and $m > N$. Then (3.6.1) becomes

$$|a_m| < \epsilon.$$

But this is precisely the conclusion that we desire.   □

EXAMPLE 3.7 The series $\sum_{j=1}^{\infty}(-1)^j$ must diverge, *even though its terms appear to be cancelling each other out.* The reason is that the summands do not tend to zero; hence the preceding proposition applies.

Write out several partial sums of this series to see more explicitly that the partial sums are $-1, +1, -1, +1, \ldots$ and hence that the series diverges.   □

We conclude this section with a necessary and sufficient condition for convergence of a series of nonnegative terms. As with some of our other results on series, it amounts to little more than a restatement of a result on sequences.

Proposition 3.8 *A series*

$$\sum_{j=1}^{\infty} a_j$$

with all $a_j \geq 0$ is convergent if and only if the sequence of partial sums is bounded.

Proof: Notice that, because the summands are nonnegative, we have

$$S_1 = a_1 \leq a_1 + a_2 = S_2\,,$$

$$S_2 = a_1 + a_2 \leq a_1 + a_2 + a_3 = S_3\,,$$

and in general

$$S_N \leq S_N + a_{N+1} = S_{N+1}\,.$$

Thus the sequence $\{S_N\}$ of partial sums forms a increasing sequence. We know that such a sequence is convergent to a finite limit if and only if it is bounded above (see Section 2.1). This completes the proof.   □

EXAMPLE 3.9 The series $\sum_{j=1}^{\infty} 1$ is divergent since the summands are nonnegative and the sequence of partial sums $\{S_N\} = \{N\}$ is unbounded.

Referring back to Example 3.3, we see that the series $\sum_{j=1}^{\infty} \frac{1}{j}$ diverges because its partial sums are unbounded.

We see from the first example that the series $\sum_{j=1}^{\infty} 2^{-j}$ converges because its partial sums are all bounded above by 1.   □

It is frequently convenient to begin a series with summation at $j = 0$ or some other term instead of $j = 1$. All of our convergence results still apply to such a series because of the Cauchy criterion. In other words, the convergence or divergence of a series will depend only on the behavior of its "tail."

Exercises

1. Discuss convergence or divergence for each of the following series:

(a) $\displaystyle\sum_{j=1}^{\infty} \frac{(2^j)^2}{j!}$

(b) $\displaystyle\sum_{j=1}^{\infty} \frac{(2j)!}{(3j)!}$

(c) $\displaystyle\sum_{j=1}^{\infty} \frac{j!}{j^j}$

(d) $\displaystyle\sum_{j=1}^{\infty} \frac{(-1)^j}{3j^2 - 5j + 6}$

(e) $\displaystyle\sum_{j=1}^{\infty} \frac{2j - 1}{3j^2 - 2}$

(f) $\displaystyle\sum_{j=1}^{\infty} \frac{2j - 1}{3j^3 - 2}$

(g) $\displaystyle\sum_{j=1}^{\infty} \frac{\log(j + 1)}{[1 + \log j]^j}$

(h) $\displaystyle\sum_{j=12}^{\infty} \frac{1}{j \log^3 j}$

(i) $\displaystyle\sum_{j=2}^{\infty} \frac{\log(2)}{\log j}$

(j) $\displaystyle\sum_{j=2}^{\infty} \frac{1}{j \log^{1.1} j}$

2. If $b_j > 0$ for every j and if $\sum_{j=1}^{\infty} b_j$ converges then prove that $\sum_{j=1}^{\infty} (b_j)^2$ converges. Prove that the assertion is false if the positivity hypothesis is omitted. How about third powers?

3. If $b_j > 0$ for every j and if $\sum_{j=1}^{\infty} b_j$ converges then prove that $\sum_{j=1}^{\infty} \frac{1}{1+b_j}$ diverges.

4. Let $\sum_{j=1}^{\infty} a_j$ be a divergent series of positive terms. Prove that there exist numbers $b_j, 0 < b_j < a_j$, such that $\sum_{j=1}^{\infty} b_j$ diverges.

Similarly, let $\sum_{j=1}^{\infty} c_j$ be a convergent series of positive terms. Prove that there exist numbers $d_j, 0 < c_j < d_j$, such that $\sum_{j=1}^{\infty} d_j$ converges.

Thus we see that there is no "smallest" divergent series and no "largest" convergent series.

5. TRUE or FALSE: If $a_j > c > 0$ and $\sum 1/a_j$ converges, then $\sum a_j$ converges.

6. If $b_j > 0$ and $\sum_j b_j$ converges then what can you say about $\sum_j b_j/(1+b_j)$?

7. If $a_j > 0$, $b_j > 0$, $\sum_j a_j^2$ converges, and $\sum_j b_j^2$ converges, then what can you say about $\sum_j a_j b_j$?

8. If $b_j > 0$ and $\sum_j b_j$ diverges, then what can you say about $\sum_j 2^{-j} b_j$?

9. If $b_j > 0$ and $\sum_j b_j$ converges, then what can you say about $\sum_j b_j/j^2$?

10. If $a_j > 0$ and $\sum_j a_j^2$ converges, then what can you say about $\sum_j a_j^4$? How about $\sum_j a_j^3$?

11. Let α and β be positive real numbers. Discuss convergence and divergence for the series

$$\sum_{j=2}^{\infty} \frac{1}{j^{\alpha} |\log j|^{\beta}} \,.$$

* **12.** Let k be a positive integer. Discuss convergence or divergence for the series

$$\sum_{j=1}^{\infty} \frac{j^k}{2^j} \,.$$

3.2 Elementary Convergence Tests

As previously noted, a series may converge because its terms are nonnegative and diminish in size fairly rapidly (thus causing its partial sums to grow slowly) or it may converge because of cancellation among the terms. The tests which measure the first type of convergence are the most obvious and these are the "elementary" ones that we discuss in the present section.

Proposition 3.10 (The Comparison Test) *Suppose that $\sum_{j=1}^{\infty} a_j$ is a convergent series of nonnegative terms. If $\{b_j\}$ are real or complex numbers and if $|b_j| \leq a_j$ for every j then the series $\sum_{j=1}^{\infty} b_j$ converges.*

Proof: Because the first series converges, its satisfies the Cauchy criterion for series. Hence, given $\epsilon > 0$, there is an N so large that if $n \geq m > N$ then

$$\left| \sum_{j=m}^{n} a_j \right| < \epsilon \,.$$

But then

$$\left| \sum_{j=m}^{n} b_j \right| \leq \sum_{j=m}^{n} |b_j| \leq \sum_{j=m}^{n} a_j < \epsilon \,.$$

It follows that the series $\sum b_j$ satisfies the Cauchy criterion for series. Therefore it converges. $\qquad \square$

Corollary 3.11 *If $\sum_{j=1}^{\infty} a_j$ is as in the proposition and if $0 \leq b_j \leq a_j$ for every j then the series $\sum_{j=1}^{\infty} b_j$ converges.*

Proof: Obvious. Simply notice that $|b_j| = b_j$. $\qquad \square$

EXAMPLE 3.12 The series $\sum_{j=1}^{\infty} 2^{-j} \sin j$ is seen to converge by comparing it with the series $\sum_{j=1}^{\infty} 2^{-j}$. $\qquad \square$

Theorem 3.13 (The Cauchy Condensation Test) *Assume that* $a_1 \geq a_2 \geq \cdots \geq a_j \geq \ldots 0$. *The series*

$$\sum_{j=1}^{\infty} a_j$$

converges if and only if the series

$$\sum_{k=1}^{\infty} 2^k \cdot a_{2^k}$$

converges.

Proof: First assume that the series $\sum_{j=1}^{\infty} a_j$ converges. Notice that, for each $k \geq 1$,

$$
\begin{aligned}
2^{k-1} \cdot a_{2^k} &= \underbrace{a_{2^k} + a_{2^k} + \cdots + a_{2^k}}_{2^{k-1} \text{ times}} \\[2mm]
&\leq \quad a_{2^{k-1}+1} + a_{2^{k-1}+2} + \cdots + a_{2^k}. \\[2mm]
&= \sum_{m=2^{k-1}+1}^{2^k} a_m
\end{aligned}
$$

Therefore

$$\sum_{k=1}^{N} 2^{k-1} \cdot a_{2^k} \leq \sum_{k=1}^{N} \sum_{m=2^{k-1}+1}^{2^k} a_m = \sum_{m=2}^{2^N} a_m.$$

Since the partial sums on the right are bounded (because the series of a_js converges), so are the partial sums on the left. It follows that the series

$$\sum_{k=1}^{\infty} 2^k \cdot a_{2^k}$$

converges.

For the converse, assume that the series

$$\sum_{k=1}^{\infty} 2^k \cdot a_{2^k} \tag{3.13.1}$$

converges. Observe that, for $k \geq 1$,

$$\sum_{m=2^{k-1}+1}^{2^k} a_j \;=\; a_{2^{k-1}+1} + a_{2^{k-1}+2} + \cdots + a_{2^k}$$

$$\leq \;\;\underbrace{a_{2^{k-1}} + a_{2^{k-1}} + \cdots + a_{2^{k-1}}}_{2^{k-1}\ \text{times}}$$

$$=\;\; 2^{k-1} \cdot a_{2^{k-1}}\,.$$

It follows that

$$\sum_{m=2}^{2^N} a_j \;=\; \sum_{k=1}^{N} \sum_{m=2^{k-1}+1}^{2^k} a_m$$

$$\leq \;\;\sum_{k=1}^{N} 2^{k-1} \cdot a_{2^{k-1}}\,.$$

By the hypothesis that the series (3.13.1) converges, the partial sums on the right must be bounded. But then the partial sums on the left are bounded as well. Since the summands a_j are nonnegative, the series on the left converges. $\square$

EXAMPLE 3.14 We apply the Cauchy condensation test to the harmonic series

$$\sum_{j=1}^{\infty} \frac{1}{j}\,.$$

It leads us to examine the series

$$\sum_{k=1}^{\infty} 2^k \cdot \frac{1}{2^k} = \sum_{k=1}^{\infty} 1\,.$$

Since the latter series diverges, the harmonic series diverges as well. $\square$

Proposition 3.15 (Geometric Series) *Let α be a complex number. The series*

$$\sum_{j=0}^{\infty} \alpha^j$$

is called a geometric series. It converges if and only if $|\alpha| < 1$. In this circumstance, the sum of the series (that is, the limit of the partial sums) is $1/(1-\alpha)$.

Proof: Let S_N denote the Nth partial sum of the geometric series. Then

$$\alpha \cdot S_N \;=\; \alpha(1 + \alpha + \alpha^2 + \ldots \alpha^N)$$

$$=\; \alpha + \alpha^2 + \ldots \alpha^{N+1}\,.$$

It follows that $\alpha \cdot S_N$ and S_N are nearly the same: in fact

$$\alpha \cdot S_N + 1 - \alpha^{N+1} = S_N \,.$$

Solving this equation for the quantity S_N yields

$$S_N = \frac{1 - \alpha^{N+1}}{1 - \alpha}$$

when $\alpha \neq 1$.

If $|\alpha| < 1$ then $\alpha^{N+1} \to 0$, hence the sequence of partial sums tends to the limit $1/(1 - \alpha)$. If $|\alpha| > 1$ then α^{N+1} diverges, hence the sequence of partial sums diverges. This completes the proof for $|\alpha| \neq 1$. But the divergence in case $|\alpha| = 1$ follows because the summands will not tend to zero. □

EXAMPLE 3.16 The series

$$\sum_{j=0}^{\infty} 3^{-j}$$

is a geometric series. Writing it as

$$\sum_{j=0}^{\infty} \left(\frac{1}{3}\right)^j ,$$

we see that the sum is

$$\frac{1}{1 - 1/3} = \frac{3}{2} \,.$$

The series

$$\sum_{j=2}^{\infty} \left(\frac{3}{4}\right)^j$$

is not quite a geometric series because the summation process does not begin at $j = 0$. But this situation is easily repaired. We write the series as

$$\left(\frac{3}{4}\right)^2 \sum_{j=0}^{\infty} \left(\frac{3}{4}\right)^j$$

and then we see that the sum is

$$\frac{9}{16} \cdot \frac{1}{1 - 3/4} = \frac{9}{16} \cdot 4 = \frac{9}{4} \,. \qquad\qquad □$$

Corollary 3.17 *Let r be a real number. The series*

$$\sum_{j=1}^{\infty} \frac{1}{j^r}$$

converges if r exceeds 1 and diverges otherwise.

Proof: When $r > 1$ we can apply the Cauchy Condensation Test. This leads us to examine the series

$$\sum_{k=1}^{\infty} 2^k \cdot 2^{-kr} = \sum_{k=1}^{\infty} \left(2^{1-r}\right)^k .$$

This last is a geometric series, with the role of α played by the quantity $\alpha = 2^{1-r}$. When $r > 1$ then $|\alpha| < 1$ so the series converges. Otherwise it diverges. $\square$

EXAMPLE 3.18 The series

$$\sum_{j=1}^{\infty} \frac{1}{j^{3/2}}$$

converges because $3/2 > 1$.

The series

$$\sum_{j=1}^{\infty} \frac{1}{j^{2/3}}$$

diverges because $2/3 < 1$. $\square$

Theorem 3.19 (The Root Test) *Consider the series*

$$\sum_{j=1}^{\infty} a_j .$$

If

$$\limsup_{j \to \infty} |a_j|^{1/j} < 1$$

then the series converges.

Proof: Refer again to the discussion of the concept of limit superior in Chapter 2. By our hypothesis, there is a number $0 < \beta < 1$ and an integer $N > 1$ such that, for all $j > N$, it holds that

$$|a_j|^{1/j} < \beta .$$

In other words,

$$|a_j| < \beta^j .$$

Since $0 < \beta < 1$ the sum of the terms on the right constitutes a convergent geometric series. By the Comparison Test, the sum of the terms on the left converges. $\square$

Theorem 3.20 (The Ratio Test) *Consider a series*

$$\sum_{j=1}^{\infty} a_j \,.$$

If

$$\limsup_{j\to\infty} \left| \frac{a_{j+1}}{a_j} \right| < 1$$

then the series converges.

Proof: It is possible to supply a proof similar to that of the Root Test. We leave such a proof for the exercises, and instead supply an argument which relates the two tests in an interesting fashion.

Let

$$\lambda = \limsup_{j\to\infty} \left| \frac{a_{j+1}}{a_j} \right| < 1 \,.$$

Select a real number μ such that $\lambda < \mu < 1$. By the definition of $\limsup$, there is an N so large that if $j > N$ then

$$\left| \frac{a_{j+1}}{a_j} \right| < \mu \,.$$

This may be rewritten as

$$|a_{j+1}| < \mu \cdot |a_j| \quad , \qquad j \geq N \,.$$

Thus (much as in the proof of the Root Test) we have for $k \geq 0$ that

$$|a_{N+k}| \leq \mu \cdot |a_{N+k-1}| \leq \mu \cdot \mu \cdot |a_{N+k-2}| \leq \cdots \leq \mu^k \cdot |a_N| \,.$$

It is convenient to denote $N + k$ by $n, n \geq N$. Thus the last inequality reads

$$|a_n| < \mu^{n-N} \cdot |a_N|$$

or

$$|a_n|^{1/n} < \mu^{(n-N)/n} \cdot |a_N|^{1/n} \,.$$

Remembering that N has been fixed once and for all, we pass to the $\limsup$ as $n \to \infty$. The result is

$$\limsup_{n\to\infty} |a_n|^{1/n} \leq \mu \,.$$

Since $\mu < 1$, we find that our series satisfies the hypotheses of the Root Test. Hence it converges.   □

Remark 3.21 The proof of the Ratio Test shows that *if* a series passes the Ratio Test then it passes the Root Test (the converse is not true, as you will learn in Exercise **2**). Put another way, the Root Test is a better test than the Ratio Test because it will give information whenever the Ratio Test does and also in some circumstances when the Ratio Test does not.

Why do we therefore learn the Ratio Test? The answer is that there are circumstances when the Ratio Test is easier to apply than the Root Test.

EXAMPLE 3.22 The series

$$\sum_{j=1}^{\infty} \frac{2^j}{j!}$$

is easily studied using the Ratio Test (recall that $j! \equiv j \cdot (j-1) \cdot \ldots 2 \cdot 1$). Indeed $a_j = 2^j/j!$ and

$$\left| \frac{a_{j+1}}{a_j} \right| = \frac{2^{j+1}/(j+1)!}{2^j/j!}.$$

We can perform the division to see that

$$\left| \frac{a_{j+1}}{a_j} \right| = \frac{2}{j+1}.$$

The lim sup of the last expression is 0. By the Ratio Test, the series converges.

Notice that in this example, while the Root Test applies in principle, it would be difficult to use in practice. $\square$

EXAMPLE 3.23 We apply the Root Test to the series

$$\sum_{j=1}^{\infty} \frac{j^2}{2^j}.$$

Observe that

$$a_j = \frac{j^2}{2^j}$$

hence that

$$|a_j|^{1/j} = \frac{\left(j^{1/j}\right)^2}{2}.$$

As $j \to \infty$, we see that

$$\limsup_{j \to \infty} |a_j|^{1/j} = \frac{1}{2}.$$

By the Root Test, the series converges. $\square$

It is natural to ask whether the Ratio and Root Tests can detect divergence. Neither test is necessary and sufficient: there are series which elude the analysis of both tests. However, the arguments that we used to establish Theorems 3.19 and 3.20 can also be used to establish the following (the proofs are left as exercises):

Theorem 3.24 (The Root Test for Divergence) *Consider the series*

$$\sum_{j=1}^{\infty} a_j$$

of nonzero terms. If

$$\liminf_{j \to \infty} |a_j|^{1/j} > 1$$

then the series diverges.

Theorem 3.25 (The Ratio Test for Divergence) *Consider the series*

$$\sum_{j=1}^{\infty} a_j \,.$$

If

$$\liminf_{j \to \infty} \left| \frac{a_{j+1}}{a_j} \right| > 1 \,,$$

then the series diverges.

In both the Root Test and the Ratio Test, if the lim sup is equal to 1, then no conclusion is possible. The exercises give examples of series, some of which converge and some of which do not, in which these tests give lim sup equal to 1.

EXAMPLE 3.26 Consider the series

$$\sum_{j=1}^{\infty} \frac{j!}{2^j} \,.$$

We apply the Ratio Test:

$$\frac{a_{j+1}}{a_j} = \frac{(j+1)!/2^{j+1}}{j!/2^j} = \frac{j+1}{2} \,.$$

This expression is > 2 for $j > 3$. Therefore, by the Ratio Test for Divergence, the series diverges. □

EXAMPLE 3.27 Consider the series

$$\sum_{j=1}^{\infty} \frac{j^j}{4^j} \,.$$

We apply the Root Test:

$$\sqrt[j]{a_j} = \sqrt[j]{j^j/4^j} = j/4 \,.$$

This expression is > 2 for $j > 8$. Therefore the lim sup is > 1 and the series diverges. □

We conclude this section by saying a word about the integral test.

Proposition 3.28 (The Integral Test) *Let f be a continuous function on $[0, \infty)$ that is monotonically decreasing. The series*

$$\sum_{j=1}^{\infty} f(j)$$

converges if and only if the integral

$$\int_1^{\infty} f(x) \, dx$$

converges.

We have not treated the integral yet in this book, so we shall not prove the result here.

EXAMPLE 3.29 Consider the harmonic series

$$\sum_{j=1}^{\infty} \frac{1}{j}.$$

The terms of this series satisfy the hypothesis of the integral test. Also

$$\int_1^{\infty} \frac{1}{x}\, dx = \lim_{N \to +\infty} \log x \Big|_1^N = \lim_{N \to +\infty} [\log N - \log 1] = +\infty.$$

Therefore the series diverges.   □

Exercises

1. Let p be a polynomial with no constant term. If $b_j > 0$ for every j and if $\sum_{j=1}^{\infty} b_j$ converges then prove that the series $\sum_{j=1}^{\infty} p(b_j)$ converges.

2. Examine the series

$$\frac{1}{3} + \frac{1}{5} + \frac{1}{3^2} + \frac{1}{5^2} + \frac{1}{3^3} + \frac{1}{5^3} + \frac{1}{3^4} + \frac{1}{5^4} + \dots$$

   Prove that the Root Test shows that the series converges while the Ratio Test gives no information.

3. Check that both the Root Test and the Ratio Test give no information for the series $\sum_{j=1}^{\infty} \frac{1}{j}$, $\sum_{j=1}^{\infty} \frac{1}{j^2}$. However, one of these series is divergent and the other is convergent.

4. Let a_j be a sequence of real numbers. Define

$$m_j = \frac{a_1 + a_2 + \dots a_j}{j}.$$

   Prove that if $\lim_{j \to \infty} a_j = \ell$ then $\lim_{j \to \infty} m_j = \ell$. Give an example to show that the converse is not true.

5. Imitate the proof of the Root Test to give a direct proof of the Ratio Test.

6. Let $\sum_j a_j$ and $\sum_j b_j$ be series of positive terms. Prove that, if there is a constant $C > 0$ such that

$$\frac{1}{C} \le \frac{a_j}{b_j} \le C$$

   for all j large, then either both series diverge or both series converge.

7. Prove that if a series of positive terms passes the Ratio Test, then it also passes the Root Test.

8. TRUE or FALSE: If the a_j are positive and $\sum a_j$ converges then $\sum a_j/j$ converges.

9. TRUE or FALSE: If a_j and b_j are positive and $\sum_j a_j$ and $\sum b_j$ both converge, then $\sum_j a_j b_j$ converges.

10. Prove Theorem 3.24.

11. Prove Theorem 3.25.

3.3   Advanced Convergence Tests

In this section we consider convergence tests for series which depend on cancellation among the terms of the series. One of the most profound of these depends on a technique called *summation by parts*. You may wonder whether this process is at all related to the "integration by parts" procedure that you learned in calculus—it has a similar form. Indeed it will turn out (and we shall see the details of this assertion as the book develops) that summing a series and performing an integration are two aspects of the same limiting process. The summation by parts method is merely our first glimpse of this relationship.

Proposition 3.30 (Summation by Parts) *Let $\{a_j\}_{j=0}^\infty$ and $\{b_j\}_{j=0}^\infty$ be two sequences of real or complex numbers. For $N = 0, 1, 2, \ldots$ set*

$$A_N = \sum_{j=0}^N a_j$$

(we adopt the convention that $A_{-1} = 0$). Then, for any $0 \le m \le n < \infty$, it holds that

$$\sum_{j=m}^n a_j \cdot b_j = [A_n \cdot b_n - A_{m-1} \cdot b_m]$$

$$+ \sum_{j=m}^{n-1} A_j \cdot (b_j - b_{j+1}).$$

Proof: We write

$$\sum_{j=m}^n a_j \cdot b_j = \sum_{j=m}^n (A_j - A_{j-1}) \cdot b_j$$

$$= \sum_{j=m}^n A_j \cdot b_j - \sum_{j=m}^n A_{j-1} \cdot b_j$$

$$= \sum_{j=m}^n A_j \cdot b_j - \sum_{j=m-1}^{n-1} A_j \cdot b_{j+1}$$

$$= \sum_{j=m}^{n-1} A_j \cdot (b_j - b_{j+1}) + A_n \cdot b_n - A_{m-1} \cdot b_m.$$

This is what we wished to prove.   □

Now we apply summation by parts to prove a convergence test due to Niels Henrik Abel (1802–1829).

Theorem 3.31 (Abel's Convergence Test) *Consider the series*

$$\sum_{j=0}^{\infty} a_j \cdot b_j .$$

Suppose that

1. *The partial sums $A_N = \sum_{j=0}^{N} a_j$ form a bounded sequence;*

2. $b_0 \geq b_1 \geq b_2 \geq \ldots$;

3. $\lim_{j \to \infty} b_j = 0.$

Then the original series

$$\sum_{j=0}^{\infty} a_j \cdot b_j$$

converges.

Proof: Suppose that the partial sums A_N are bounded in absolute value by a number K. Pick $\epsilon > 0$ and choose an integer N so large that $b_N < \epsilon/(2K)$. For $N < m \leq n < \infty$ we use the partial summation formula to write

$$\left| \sum_{j=m}^{n} a_j \cdot b_j \right| = \left| A_n \cdot b_n - A_{m-1} \cdot b_m + \sum_{j=m}^{n-1} A_j \cdot (b_j - b_{j+1}) \right|$$

$$\leq K \cdot |b_n| + K \cdot |b_m| + K \cdot \sum_{j=m}^{n-1} |b_j - b_{j+1}| .$$

Now we take advantage of the facts that $b_j \geq 0$ for all j and that $b_j \geq b_{j+1}$ for all j to estimate the last expression by

$$K \cdot \left[b_n + b_m + \sum_{j=m}^{n-1} (b_j - b_{j+1}) \right] .$$

[Notice that the expressions $b_j - b_{j+1}, b_m$, and b_n are all nonnegative.] Now the sum collapses and the last line is estimated by

$$K \cdot [b_n + b_m - b_n + b_m] = 2 \cdot K \cdot b_m .$$

By our choice of N the right side is smaller than ϵ. Thus our series satisfies the Cauchy criterion and therefore converges.   □

EXAMPLE 3.32 (THE ALTERNATING SERIES TEST) As a first application of
Abel's convergence test, we examine alternating series. Consider a series of the
form

$$\sum_{j=1}^{\infty}(-1)^j \cdot b_j, \qquad\qquad (3.32.1)$$

with $b_1 \geq b_2 \geq b_3 \geq \cdots \geq 0$ and $b_j \to 0$ as $j \to \infty$. We set $a_j = (-1)^j$ and apply
Abel's test. We see immediately that all partial sums A_N are either -1 or 0. In
particular, this sequence of partial sums is bounded. And the b_js are decreasing
and tending to zero. By Abel's convergence test, the alternating series (3.32.1)
converges.   □

Proposition 3.33 *Let $b_1 \geq b_2 \geq \ldots$ and assume that $b_j \to 0$. Consider the
alternating series $\sum_{j=1}^{\infty}(-1)^j b_j$ as in the last example. It is convergent: let S
be its sum. Then the partial sums S_N satisfy $|S - S_N| \leq b_{N+1}$.*

Proof: Observe that

$$|S - S_N| = |b_{N+1} - b_{N+2} + b_{N+3} - + \ldots|.$$

But

$$
\begin{aligned}
b_{N+2} - b_{N+3} + - \ldots \quad &\leq \quad b_{N+2} + (-b_{N+3} + b_{N+3}) \\
&\qquad\qquad + (-b_{N+5} + b_{N+5}) + \ldots \\
&= \quad b_{N+2}
\end{aligned}
$$

and

$$
\begin{aligned}
b_{N+2} - b_{N+3} + - \ldots \quad &\geq \quad (b_{N+2} - b_{N+2}) + (b_{N+4} - b_{N+4}) + \ldots \\
&= \quad 0.
\end{aligned}
$$

It follows that

$$|S - S_N| \leq |b_{N+1}|$$

as claimed.   □

EXAMPLE 3.34 Consider the series

$$\sum_{j=1}^{\infty}(-1)^j \frac{1}{j}.$$

Then the partial sum $S_{100} = -.688172$ is within 0.01 (in fact within $1/101$) of
the full sum S and the partial sum $S_{10000} = -.6930501$ is within 0.0001 (in fact
within $1/10001$) of the sum S.   □

EXAMPLE 3.35 Next we examine a series which is important in the study of Fourier analysis. Consider the series

$$\sum_{j=1}^{\infty} \frac{\sin j}{j} . \tag{3.35.1}$$

We already know that the series $\sum \frac{1}{j}$ diverges. However, the expression $\sin j$ changes sign in a rather sporadic fashion. We might hope that the series (3.35.1) converges because of cancellation of the summands. We take $a_j = \sin j$ and $b_j = 1/j$. Abel's test will apply if we can verify that the partial sums A_N of the a_js are bounded. To see this we use a trick:

Observe that

$$\cos(j + 1/2) = \cos j \cdot \cos 1/2 - \sin j \cdot \sin 1/2$$

and

$$\cos(j - 1/2) = \cos j \cdot \cos 1/2 + \sin j \cdot \sin 1/2 .$$

Subtracting these equations and solving for $\sin j$ yields that

$$\sin j = \frac{\cos(j - 1/2) - \cos(j + 1/2)}{2 \cdot \sin 1/2} .$$

We conclude that

$$A_N = \sum_{j=1}^{N} a_j = \sum_{j=1}^{N} \frac{\cos(j - 1/2) - \cos(j + 1/2)}{2 \cdot \sin 1/2} .$$

Of course this sum collapses and we see that

$$A_N = \frac{-\cos(N + 1/2) + \cos 1/2}{2 \cdot \sin 1/2} .$$

Thus

$$|A_N| \leq \frac{2}{2 \cdot \sin 1/2} = \frac{1}{\sin 1/2} ,$$

independent of N.

Thus the hypotheses of Abel's test are verified and the series

$$\sum_{j=1}^{\infty} \frac{\sin j}{j}$$

converges. $\square$

Remark 3.36 It is interesting to notice that both the series

$$\sum_{j=1}^{\infty} \frac{|\sin j|}{j} \quad \text{and} \quad \sum_{j=1}^{\infty} \frac{\sin^2 j}{j}$$

diverge. The proofs of these assertions are left as exercises for you.

We turn next to the topic of absolute and conditional convergence.

Definition 3.37 A series of real or complex numbers

$$\sum_{j=1}^{\infty} a_j$$

is said to be *absolutely convergent* if

$$\sum_{j=1}^{\infty} |a_j|$$

converges.

We have:

Proposition 3.38 *If the series $\sum_{j=1}^{\infty} a_j$ is absolutely convergent, then it is convergent.*

Proof: This is an immediate corollary of the Comparison Test.   □

Definition 3.39 A series $\sum_{j=1}^{\infty} a_j$ is said to be *conditionally convergent* if $\sum_{j=1}^{\infty} a_j$ converges, but it does not converge absolutely.

We see that absolutely convergent series are convergent but the next example shows that the converse is not true.

EXAMPLE 3.40 The series

$$\sum_{j=1}^{\infty} \frac{(-1)^j}{j}$$

converges by the Alternating Series Test. However, it is not absolutely convergent because the harmonic series

$$\sum_{j=1}^{\infty} \frac{1}{j}$$

diverges.   □

There is a remarkable robustness result for absolutely convergent series that fails dramatically for conditionally convergent series. This result is enunciated in the next theorem. We first need a definition.

Definition 3.41 Let $\sum_{j=1}^{\infty} a_j$ be a given series. Let $\{p_j\}_{j=1}^{\infty}$ be a sequence in which every positive integer occurs once and only once (but not necessarily in the usual order). We call $\{p_j\}$ a *permutation* of the natural numbers.

Then the series

$$\sum_{j=1}^{\infty} a_{p_j}$$

is said to be a *rearrangement* of the given series.

Theorem 3.42 (Riemann, Weierstrass) *If the series $\sum_{j=1}^{\infty} a_j$ of real numbers is absolutely convergent and if the sum of the series is ℓ, then every rearrangement of the series converges also to ℓ.*

If the real series $\sum_{j=1}^{\infty} b_j$ is conditionally convergent and if β is any real number or $\pm\infty$ then there is a rearrangement of the series that converges to β.

Proof: We prove the first assertion here and explore the second in the exercises.

Let us choose a rearrangement of the given series and denote it by $\sum_{j=1}^{\infty} a_{p_j}$, where p_j is a permutation of the positive integers. Pick $\epsilon > 0$. By the hypothesis that the original series converges absolutely we may choose an integer $N > 0$ such that $N < m \leq n < \infty$ implies that

$$\sum_{j=m}^{n} |a_j| < \epsilon. \qquad (3.42.1)$$

[The presence of the absolute values in the left side of this inequality will prove crucial in a moment.] Choose a positive integer M such that $M \geq N$ and the integers $1, \ldots, N$ are all contained in the list $p_1, p_2, \ldots, p_M$. If $K > M$ then the partial sum $\sum_{j=1}^{K} a_j$ will trivially contain the summands $a_1, a_2, \ldots a_N$. Also the partial sum $\sum_{j=1}^{K} a_{p_j}$ will contain the summands $a_1, a_2, \ldots a_N$. It follows that

$$\sum_{j=1}^{K} a_j - \sum_{j=1}^{K} a_{p_j}$$

will contain only summands *after* the Nth one in the original series. By inequality (3.42.1) we may conclude that

$$\left| \sum_{j=1}^{K} a_j - \sum_{j=1}^{K} a_{p_j} \right| \leq \sum_{j=N+1}^{\infty} |a_j| \leq \epsilon.$$

We conclude that the rearranged series converges; and it converges to the same sum as the original series. $\square$

Exercises

1. If $1/2 > b_j > 0$ for every j and if $\sum_{j=1}^{\infty} b_j$ converges then prove that $\sum_{j=1}^{\infty} \frac{b_j}{1-b_j}$ converges.

2. Follow these steps to give another proof of the Alternating Series Test: **a)** Prove that the odd partial sums form an increasing sequence; **b)** Prove that the even partial sums form a decreasing sequence; **c)** Prove that every even partial sum majorizes all subsequent odd partial sums; **d)** Use a pinching principle.

3. What can you say about the convergence or divergence of

$$\sum_{j=1}^{\infty} \frac{(2j+3)^{1/2} - (2j)^{1/2}}{j^{3/4}} ?$$

4. For which exponents k and ℓ does the series

$$\sum_{j=2}^{\infty} \frac{1}{j^k |\log j|^{\ell}}$$

converge?

5. Let p be a polynomial with integer coefficients and degree at least 1. Let $b_1 \geq b_2 \geq \cdots \geq 0$ and assume that $b_j \to 0$. Prove that if $(-1)^{p(j)}$ is not always positive and not always negative then in fact it will alternate in sign so that $\sum_{j=1}^{\infty} (-1)^{p(j)} \cdot b_j$ will converge.

6. Explain in words how summation by parts is analogous to integration by parts.

7. If $\gamma_j > 0$ and $\sum_{j=1}^{\infty} \gamma_j$ converges then prove that

$$\sum_{j=1}^{\infty} (\gamma_j)^{1/2} \cdot \frac{1}{j^{\alpha}}$$

converges for any $\alpha > 1/2$. Give an example to show that the assertion is false if $\alpha = 1/2$.

* **8.** Assume that $\sum_{j=1}^{\infty} b_j$ is a convergent series of positive real numbers. Let $s_j = \sum_{\ell=1}^{j} b_\ell$. Discuss convergence or divergence for the series $\sum_{j=1}^{\infty} s_j \cdot b_j$. Discuss convergence or divergence for the series $\sum_{j=1}^{\infty} \frac{b_j}{1+s_j}$.

* **9.** If $b_j > 0$ for every j and if $\sum_{j=1}^{\infty} b_j$ diverges then define $s_j = \sum_{\ell=1}^{j} b_\ell$. Discuss convergence or divergence for the series $\sum_{j=1}^{\infty} \frac{b_j}{s_j}$.

* **10.** Let $\sum_{j=1}^{\infty} b_j$ be a conditionally convergent series of real numbers. Let β be a real number. Prove that there is a rearrangement of the series that converges to β. (**Hint:** First observe that the positive terms of the given series must form a divergent series. Also, the negative terms form a divergent series. Now build the rearrangement by choosing finitely many positive terms whose sum "just exceeds" β. Then add on enough negative terms so that the sum is "just less than" β. Repeat this oscillatory procedure.)

* **11.** Do Exercise 10 in the case that β is $\pm\infty$.

* **12.** Let $\sum_{j=1}^{\infty} a_j$ be a conditionally convergent series of complex numbers. Let $\mathcal{S}$ be the set of all possible complex numbers to which the various rearrangements could converge. What forms can $\mathcal{S}$ have? (**Hint:** Experiment!)

3.4   Some Special Series

We begin with a series that defines a special constant of mathematical analysis.

Definition 3.43 The series

$$\sum_{j=0}^{\infty} \frac{1}{j!},$$

where $j! \equiv j \cdot (j-1) \cdot (j-2) \cdots 1$ for $j \geq 1$ and $0! \equiv 1$, is convergent (by the Ratio Test, for instance). Its sum is denoted by the symbol e in honor of the Swiss mathematician Léonard Euler, who first studied it (see also Example 2.47, where the number e is studied by way of a sequence). We shall see in Proposition 3.44 that these two approaches to the number e are equivalent.

Like the number π, to be considered later in this book, the number e is one which arises repeatedly in a number of contexts in mathematics. It has many special properties. We first relate the series definition of e to the sequence definition:

Proposition 3.44 *The limit*

$$\lim_{n \to \infty} \left(1 + \frac{1}{n}\right)^n$$

exists and equals e.

Proof: We need to compare the quantities

$$A_N \equiv \sum_{j=0}^{N} \frac{1}{j!} \quad \text{and} \quad B_N \equiv \left(1 + \frac{1}{N}\right)^N.$$

We use the binomial theorem to expand B_N :

$$\begin{aligned}
B_N &= 1 + \frac{N}{1} \cdot \frac{1}{N} + \frac{N \cdot (N-1)}{2 \cdot 1} \cdot \frac{1}{N^2} + \frac{N \cdot (N-1) \cdot (N-2)}{3 \cdot 2 \cdot 1} \cdot \frac{1}{N^3} \\
&\quad + \cdots + \frac{N}{1} \cdot \frac{1}{N^{N-1}} + 1 \cdot \frac{1}{N^N} \\
&= 1 + 1 + \frac{1}{2!} \cdot \frac{N-1}{N} + \frac{1}{3!} \cdot \frac{N-1}{N} \cdot \frac{N-2}{N} + \cdots \\
&\quad + \frac{1}{(N-1)!} \cdot \frac{N-1}{N} \cdot \frac{N-2}{N} \cdots \frac{2}{N} \\
&\quad + \frac{1}{N!} \cdot \frac{N-1}{N} \cdot \frac{N-2}{N} \cdots \frac{1}{N} \\
&= 1 + 1 + \frac{1}{2!} \cdot \left(1 - \frac{1}{N}\right) + \frac{1}{3!} \cdot \left(1 - \frac{1}{N}\right) \cdot \left(1 - \frac{2}{N}\right) + \cdots \\
&\quad + \frac{1}{(N-1)!} \cdot \left(1 - \frac{1}{N}\right) \cdot \left(1 - \frac{2}{N}\right) \cdots \left(1 - \frac{N-2}{N}\right) \\
&\quad + \frac{1}{N!} \cdot \left(1 - \frac{1}{N}\right) \cdot \left(1 - \frac{2}{N}\right) \cdots \left(1 - \frac{N-1}{N}\right).
\end{aligned}$$

Notice that every summand that appears in this last equation is positive. Thus, for $0 \leq M \leq N$,

$$
\begin{aligned}
B_N \quad \geq \quad & 1 + 1 + \frac{1}{2!} \cdot \left(1 - \frac{1}{N}\right) + \frac{1}{3!} \cdot \left(1 - \frac{1}{N}\right) \cdot \left(1 - \frac{2}{N}\right) \\
& + \cdots + \frac{1}{M!} \cdot \left(1 - \frac{1}{N}\right) \left(1 - \frac{2}{N}\right) \cdots \left(1 - \frac{M-1}{N}\right) .
\end{aligned}
$$

In this last inequality we hold M fixed and let N tend to infinity. The result is that

$$
\liminf_{N \to \infty} B_N \geq 1 + 1 + \frac{1}{2!} + \frac{1}{3!} + \cdots + \frac{1}{M!} = A_M .
$$

Now, as $M \to \infty$, the quantity A_M converges to e (by the *definition* of e). So we obtain

$$
\liminf_{N \to \infty} B_N \geq e . \tag{3.44.1}
$$

On the other hand, our expansion for B_N allows us to observe that $B_N \leq A_N$. Thus

$$
\limsup_{N \to \infty} B_N \leq e . \tag{3.44.2}
$$

Combining (3.44.1) and (3.44.2) we find that

$$
e \leq \liminf_{N \to \infty} B_N \leq \limsup_{N \to \infty} B_N \leq e
$$

hence that $\lim_{N \to \infty} B_N$ exists and equals e. This is the desired result. $\square$

Remark 3.45 The last proof illustrates the value of the concepts of $\liminf$ and $\limsup$. For we do not know in advance that the limit of the expressions B_N exists, much less that the limit equals e. However, the $\liminf$ and the $\limsup$ always exist. So we estimate those instead, and find that they are equal and that they equal e.

The next result tells us how rapidly the partial sums A_N of the series defining e converge to e. This is of theoretical interest, but will also be applied to determine the irrationality of e.

Proposition 3.46 With A_N as above, we have that

$$
0 < e - A_N < \frac{1}{N \cdot N!} .
$$

Proof: Observe that

$$
\begin{aligned}
e - A_N \quad = \quad & \frac{1}{(N+1)!} + \frac{1}{(N+2)!} + \frac{1}{(N+3)!} + \cdots \\
= \quad & \frac{1}{(N+1)!} \cdot \left(1 + \frac{1}{N+2} + \frac{1}{(N+2)(N+3)} + \cdots\right) \\
< \quad & \frac{1}{(N+1)!} \cdot \left(1 + \frac{1}{N+1} + \frac{1}{(N+1)^2} + \cdots\right) .
\end{aligned}
$$

Now the expression in parentheses is a geometric series. It sums to $(N+1)/N$. Since $A_N < e$, we have

$$e - A_N = |e - A_N|$$

hence

$$|e - A_N| < \frac{1}{N \cdot N!} \, ,$$

proving the result. $\qquad\square$

Next we prove that e is an irrational number.

Theorem 3.47 *Euler's number e is irrational.*

Proof: Suppose to the contrary that e is rational. Then $e = p/q$ for some positive integers p and q. By the preceding proposition,

$$0 < e - A_q < \frac{1}{q \cdot q!}$$

or

$$0 < q! \cdot (e - A_q) < \frac{1}{q} \, . \tag{3.47.1}$$

Now

$$e - A_q = \frac{p}{q} - \left(1 + 1 + \frac{1}{2!} + \frac{1}{3!} + \cdots + \frac{1}{q!} \right)$$

hence

$$q! \cdot (e - A_q)$$

is an integer. But then equation (3.47.1) says that this integer lies between 0 and $1/q$. In particular, this integer lies strictly between 0 and 1. That, of course, is impossible. So e must be irrational. $\qquad\square$

It is a general principle of number theory that a real number that can be approximated *too rapidly* by rational numbers (the degree of rapidity being measured in terms of powers of the denominators of the rational numbers) must be irrational. Under suitable conditions an even stronger conclusion holds: namely, the number in question turns out to be *transcendental*. A transcendental number is one which is not the solution of any polynomial equation with integer coefficients.

The subject of transcendental numbers is explored in the exercises. The exercises also contain a sketch of a proof that e is transcendental.

In Exercise **5** of Section 2.4, we briefly discussed Euler's number γ. Both this special number and also the more commonly encountered number π arise in many contexts in mathematics. It is unknown whether γ is rational or irrational. The number π is known to be transcendental, but it is unknown whether $\pi + e$ (where e is Euler's number) is transcendental.

In recent years, questions about the the irrationality and transcendence of various numbers have become a matter of practical interest. For these properties prove to be useful in making and breaking secret codes, and in encrypting information so that it is accessible to some users but not to others.

In Example A1.1 we prove that

$$S_N \equiv \sum_{j=1}^{N} j = \frac{N \cdot (N+1)}{2} \,.$$

We conclude this section with a method for summing higher powers of j.

Suppose that we wish to calculate

$$S_{k,N} \equiv \sum_{j=1}^{N} j^k$$

for some positive integer k exceeding 1. We may proceed as follows: Write

$$
\begin{aligned}
(j+1)^{k+1} - j^{k+1} &= \left[j^{k+1} + (k+1) \cdot j^k + \frac{(k+1) \cdot k}{2} \cdot j^{k-1} \right. \\
&\quad + \cdots + \frac{(k+1) \cdot k}{2} \cdot j^2 + (k+1) \cdot j + 1 \Big] \\
&\quad - j^{k+1} \\
&= (k+1) \cdot j^k + \frac{(k+1) \cdot k}{2} \cdot j^{k-1} + \cdots \\
&\quad + \frac{(k+1) \cdot k}{2} \cdot j^2 + (k+1) \cdot j + 1 \,.
\end{aligned}
$$

Summing from $j = 1$ to $j = N$ yields

$$
\begin{aligned}
\sum_{j=1}^{N} \{ (j+1)^{k+1} - j^{k+1} \} &= (k+1) \cdot S_{k,N} + \frac{(k+1) \cdot k}{2} \cdot S_{k-1,N} + \cdots \\
&\quad + \frac{(k+1) \cdot k}{2} \cdot S_{2,N} + (k+1) \cdot S_{1,N} + N \,.
\end{aligned}
$$

The sum on the left collapses to $(N+1)^{k+1} - 1$. We may solve for $S_{k,N}$ and obtain

$$
\begin{aligned}
S_{k,N} &= \frac{1}{k+1} \cdot \left[(N+1)^{k+1} - 1 - N - \frac{(k+1) \cdot k}{2} \cdot S_{k-1,N} \right. \\
&\quad - \cdots \quad \left. - \frac{(k+1) \cdot k}{2} \cdot S_{2,N} - (k+1) \cdot S_{1,N} \right] \,.
\end{aligned}
$$

We have succeeded in expressing $S_{k,N}$ in terms of $S_{1,N}, S_{2,N}, \ldots, S_{k-1,N}$. Thus we may inductively obtain formulas for $S_{k,N}$, any k. It turns out that

$$S_{1,N} = \frac{N(N+1)}{2}$$

$$S_{2,N} = \frac{N(N+1)(2N+1)}{6}$$

$$S_{3,N} = \frac{N^2(N+1)^2}{4}$$

$$S_{4,N} = \frac{(N+1)N(2N+1)(3N^2+3N-1)}{30}$$

These formulas are treated in further detail in the exercises.

Exercises

1. Use induction to prove the formulas provided in the text for the sum of the first N perfect squares, the first N perfect cubes, and the first N perfect fourth powers.

2. A real number s is called *algebraic* if it satisfies a polynomial equation of the form

$$a_0 + a_1x + a_2x^2 + \cdots + a_mx^m = 0$$

with the coefficients a_j being integers and $a_m \neq 0$. Prove that, if we replace the word "integers" in this definition with "rational numbers," then the set of algebraic numbers remains the same. Prove that $n^{p/q}$ is algebraic for any positive integers p, q, n.

A number which is not algebraic is called *transcendental*.

3. Discuss convergence of $\sum_j 1/[\ln j]^k$ for k a positive integer.

4. Discuss convergence of $\sum_j 1/p(j)$ for p a polynomial.

5. Discuss convergence of $\sum_j \exp(p(j))$ for p a polynomial.

* 6. Refer to Exercise **2** for terminology. Prove that the sum or difference of two algebraic numbers is algebraic.

7. Refer to Exercise **6**. It is not known whether $\pi+e$ or $\pi-e$ is transcendental. But one of them must be. Explain.

* 8. Refer to Exercise **2** for terminology. A real number is called *transcendental* if it is not algebraic. Prove that the number of algebraic numbers is countable. Explain why this implies that the number of transcendental numbers is uncountable. Thus most real numbers are transcendental; however, it is extremely difficult to verify that any particular real number is transcendental.

*   **9.** Refer to Exercise **2** for terminology. Provide the details of the following sketch of a proof that Euler's number e is transcendental. [Note: in this argument we use some simple ideas of calculus. These ideas will be treated in rigorous detail later in the book.] Seeking a contradiction, we suppose that the number e satisfies a polynomial equation of the form

$$a_0 + a_1 x + \cdots + a_m x^m = 0$$

with integer coefficients a_j.

   (a) We may assume that $a_0 \neq 0$.

   (b) Let p be an odd prime that will be specified later. Define

$$g(x) = \frac{x^{p-1}(x-1)^p \cdots (x-m)^p}{(p-1)!}$$

   and

$$G(x) = g(x) + g^{(1)}(x) + g^{(2)}(x) + \cdots g^{(mp+p-1)}(x).$$

   (Here parenthetical exponents denote derivatives.) Verify that

$$|g(x)| < \frac{m^{mp+p-1}}{(p-1)!}$$

   for a suitable range of x.

   (c) Check that

$$\frac{d}{dx}\left\{ e^{-x} G(x) \right\} = -e^{-x} g(x)$$

   and thus that

$$a_j \int_0^j e^{-x} g(x) dx = a_j G(0) - a_j e^{-j} G(j). \qquad (*)$$

   (d) Multiply the last equation by e^j, sum from $j = 0$ to $j = m$, and use the polynomial equation that e satisfies to obtain that

$$\sum_{j=0}^m a_j e^j \int_0^j e^{-x} g(x) dx = -\sum_{j=0}^m \sum_{i=0}^{mp+p-1} a_j g^{(i)}(j). \qquad (**)$$

   (e) Check that $g^{(i)}(j)$ is an integer for all values of i and all j from 0 to m inclusive.

   (f) Referring to the last step, show that in fact $g^{(i)}(j)$ is an integer divisible by p *except* in the case that $j = 0$ and $i = p - 1$.

(g) Check that

$$g^{(p-1)}(0) = (-1)^p(-2)^p \cdots (-m)^p.$$

Conclude that $g^{(p-1)}(0)$ is not divisible by p if $p > m$.

(h) Check that if $p > |a_0|$ then the right side of equation $(**)$ consists of a sum of terms each of which is a multiple of p *except* for the term $-a_0 g^{(p-1)}(0)$. It follows that the sum on the right side of $(**)$ is a nonzero integer.

(i) Use equation $(*)$ to check that, provided p is chosen sufficiently large, the left side of $(**)$ satisfies

$$\left| \sum_{j=0}^{m} a_j e^j \int_0^j e^{-x} g(x) dx \right| \leq \left\{ \sum_{j=0}^{m} |a_j| \right\} e^m \frac{(m^{m+2})^{p-1}}{(p-1)!} < 1.$$

(j) The last two steps contradict each other.

This proof is from [NIV].

* **10.** What can you say about the convergence of $\sum_j [\sin j]^k / j$ for k a positive integer?

3.5   Operations on Series

Some operations on series, such as addition, subtraction, and scalar multiplication, are straightforward. Others, such as multiplication, entail subtleties. This section treats all these matters.

Proposition 3.48 *Let*

$$\sum_{j=1}^{\infty} a_j \quad and \quad \sum_{j=1}^{\infty} b_j$$

be convergent series of real or complex numbers; assume that the series sum to limits α and β respectively. Then

(a) *The series $\sum_{j=1}^{\infty}(a_j + b_j)$ converges to the limit $\alpha + \beta$.*

(b) *If c is a constant then the series $\sum_{j=1}^{\infty} c \cdot a_j$ converges to $c \cdot \alpha$.*

Proof: We shall prove assertion **(a)** and leave the easier assertion **(b)** as an exercise.

Pick $\epsilon > 0$. Choose an integer N_1 so large that $n > N_1$ implies that the partial sum $S_n \equiv \sum_{j=1}^{n} a_j$ satisfies $|S_n - \alpha| < \epsilon/2$. Choose N_2 so large that $n > N_2$ implies that the partial sum $T_n \equiv \sum_{j=1}^{n} b_j$ satisfies $|T_n - \beta| < \epsilon/2$. If U_n

is the nth partial sum of the series $\sum_{j=1}^{\infty}(a_j + b_j)$ and if $n > N_0 \equiv \max(N_1, N_2)$ then

$$|U_n - (\alpha + \beta)| \leq |S_n - \alpha| + |T_n - \beta| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon\,.$$

Thus the sequence $\{U_n\}$ converges to $\alpha + \beta$. This proves part **(a)**. The proof of **(b)** is similar. $\qquad\square$

In order to keep our discussion of multiplication of series as straightforward as possible, we deal at first with absolutely convergent series. It is convenient in this discussion to begin our sums at $j = 0$ instead of $j = 1$. If we wish to multiply

$$\sum_{j=0}^{\infty} a_j \quad \text{and} \quad \sum_{j=0}^{\infty} b_j\,,$$

then we need to specify what the partial sums of the product series should be. An obvious necessary condition that we wish to impose is that, if the first series converges to α and the second converges to β, then the product series, whatever we define it to be, should converge to $\alpha \cdot \beta$.

The naive method for defining the summands of the product series $\sum_j c_j$ is to let $c_j = a_j \cdot b_j$. However, a glance at the product of two partial sums of the given series shows that such a definition would be ignoring the distributivity of multiplication over addition.

Cauchy's idea was that the summands for the product series should be

$$c_m \equiv \sum_{j=0}^{m} a_j \cdot b_{m-j}\,.$$

This particular form for the summands can be easily motivated using power series considerations (which we shall provide in Section 9.1). For now we concentrate on verifying that this "Cauchy product" of two series really works.

Theorem 3.49 *Let $\sum_{j=0}^{\infty} a_j$ and $\sum_{j=0}^{\infty} b_j$ be two absolutely convergent series which converge to limits α and β respectively. Define the series $\sum_{m=0}^{\infty} c_m$ with summands $c_m = \sum_{j=0}^{m} a_j \cdot b_{m-j}$. Then the series $\sum_{m=0}^{\infty} c_m$ converges absolutely to $\alpha \cdot \beta$.*

Proof: Let A_n, B_n, and C_n be the partial sums of the three series in question. We calculate that

$$\begin{aligned} C_n &= (a_0 b_0) + (a_0 b_1 + a_1 b_0) + (a_0 b_2 + a_1 b_1 + a_2 b_0) \\ &\quad + \cdots + (a_0 b_n + a_1 b_{n-1} + \cdots + a_n b_0) \\ &= a_0 \cdot B_n + a_1 \cdot B_{n-1} + a_2 \cdot B_{n-2} + \cdots + a_n \cdot B_0\,. \end{aligned}$$

We set $\lambda_n = B_n - \beta$, each n, and rewrite the last line as

$$\begin{aligned} C_n &= a_0(\beta + \lambda_n) + a_1(\beta + \lambda_{n-1}) + \cdots a_n(\beta + \lambda_0) \\ &= A_n \cdot \beta + [a_0 \lambda_n + a_1 \cdot \lambda_{n-1} + \cdots + a_n \cdot \lambda_0]\,. \end{aligned}$$

Denote the expression in square brackets by the symbol ρ_n. Suppose that we could show that $\lim_{n \to \infty} \rho_n = 0$. Then we would have

$$
\begin{aligned}
\lim_{n \to \infty} C_n &= \lim_{n \to \infty} (A_n \cdot \beta + \rho_n) \\
&= (\lim_{n \to \infty} A_n) \cdot \beta + (\lim_{n \to \infty} \rho_n) \\
&= \alpha \cdot \beta + 0 \\
&= \alpha \cdot \beta \,.
\end{aligned}
$$

Thus it is enough to examine the limit of the expressions ρ_n.

Since $\sum_{j=1}^{\infty} a_j$ is absolutely convergent, we know that $A = \sum_{j=1}^{\infty} |a_j|$ is a finite number. Choose $\epsilon > 0$. Since $\sum_{j=1}^{\infty} b_j$ converges to β it follows that $\lambda_n \to 0$. Thus we may choose an integer $N > 0$ such that $n > N$ implies that $|\lambda_n| < \epsilon$. Thus, for $n = N + k, k > 0$, we may estimate

$$
\begin{aligned}
|\rho_{N+k}| &\leq |\lambda_0 a_{N+k} + \lambda_1 a_{N+k-1} + + \cdots + \lambda_N a_k| \\
&\quad + |\lambda_{N+1} a_{k-1} + \lambda_{N+2} a_{k-2} + \cdots + \lambda_{N+k} a_0| \\
&\leq |\lambda_0 a_{N+k} + \lambda_1 a_{N+k-1} + + \cdots + \lambda_N a_k| \\
&\quad + \max_{p \geq 1} \{|\lambda_{N+p}|\} \cdot (|a_{k-1}| + |a_{k-2}| + \cdots + |a_0|) \\
&\leq (N+1) \cdot \max_{\ell \geq k} |a_\ell| \cdot \max_{0 \leq j \leq N} |\lambda_j| + \epsilon \cdot A \,.
\end{aligned}
$$

In this last estimate, we have used the fact (for the first term in absolute values) that **(a)** there are $N + 1$ summands, **(b)** the a terms all have index at least k, and **(c)** the λ terms have index between 0 and N. The second term (the "max" term) is easy to estimate because of our bound on λ_n.

With N fixed, we let $k \to \infty$ in the last inequality. Since $\max_{\ell \geq k} |a_\ell| \to 0$, we find that

$$
\limsup_{n \to \infty} |\rho_n| \leq \epsilon \cdot A \,.
$$

Since $\epsilon > 0$ was arbitrary, we conclude that

$$
\lim_{n \to \infty} |\rho_n| \to 0 \,.
$$

This completes the proof. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad\square$

Notice that, in the proof of the theorem, we really only used the fact that one of the given series was absolutely convergent, not that both were absolutely convergent. Some hypothesis of this nature is necessary, as the following example shows.

EXAMPLE 3.50 Consider the Cauchy product of the two conditionally convergent series

$$
\sum_{j=0}^{\infty} \frac{(-1)^j}{\sqrt{j+1}} \quad \text{and} \quad \sum_{j=0}^{\infty} \frac{(-1)^j}{\sqrt{j+1}} \,.
$$

Observe that

$$c_m = \frac{(-1)^0(-1)^m}{\sqrt{1}\sqrt{m+1}} + \frac{(-1)^1(-1)^{m-1}}{\sqrt{2}\sqrt{m}} + \cdots$$
$$+ \frac{(-1)^m(-1)^0}{\sqrt{m+1}\sqrt{1}}$$
$$= \sum_{j=0}^{m} (-1)^m \frac{1}{\sqrt{(j+1)\cdot(m+1-j)}}.$$

However, for $0 \leq j \leq m$,

$$(j+1)\cdot(m+1-j) \leq (m+1)\cdot(m+1) = (m+1)^2.$$

Thus

$$|c_m| \geq \sum_{j=0}^{m} \frac{1}{m+1} = 1.$$

We thus see that the terms of the series $\sum_{m=0}^{\infty} c_m$ do not tend to zero, so the series cannot converge. $\qquad\square$

Exercises

1. Calculate the Cauchy product of the series $\sum_j 1/j^3$ and the series $\sum_j 1/j^4$.

2. Explain how you could discover the Cauchy product using multiplication of polynomials.

3. Discuss the concept of composition of power series.

4. Let $\sum_{j=1}^{\infty} a_j$ and $\sum_{j=1}^{\infty} b_j$ be convergent series of positive real numbers. Discuss division of these two series. Use the idea of the Cauchy product.

5. Let $\sum_{j=1}^{\infty} a_j$ and $\sum_{j=1}^{\infty} b_j$ be convergent series of positive real numbers. Discuss convergence of $\sum_{j=1}^{\infty} a_j b_j$.

6. If $\sum_j a_j$ is a convergent series of positive terms and if $\sum_j b_j$ is a convergent series of positive terms, then what can you say about $\sum_j (a_j/b_j)$?

7. Prove Proposition 3.48(b).

* 8. Explain division of power series in the language of the Cauchy product.

* 9. Discuss the concept of the exponential of a power series.

* 10. Is there a way to calculate the square root of a power series?

* 11. Is there a way to calculate the logarithm of a power series?

Chapter 4

Basic Topology

4.1 Open and Closed Sets

To specify a topology on a set is to describe certain subsets that will play the role of neighborhoods. These sets are called *open sets*. Our purpose here is to be able to study the "shape" of a set without worrying about its rigid properties. People like to joke that a mathematician does not know a coffee cup from a donut because they both have the same shape—a loop with a hole in the middle. See Figure 4.1. The purpose of this chapter is to make these ideas precise.

In what follows, we will use "interval notation": If $a \leq b$ are real numbers then we define

$$
\begin{aligned}
(a, b) &= \{x \in \mathbb{R} : a < x < b\}, \\
[a, b] &= \{x \in \mathbb{R} : a \leq x \leq b\}, \\
[a, b) &= \{x \in \mathbb{R} : a \leq x < b\}, \\
(a, b] &= \{x \in \mathbb{R} : a < x \leq b\}.
\end{aligned}
$$

Intervals of the form (a, b) are called *open*. Those of the form $[a, b]$ are called *closed*. The other two are termed *half-open* or *half-closed*. See Figure 4.2.

Now we extend the terms "open" and "closed" to more general sets.

Definition 4.1 A set $U \subseteq \mathbb{R}$ is called *open* if, for each $x \in U$, there is an $\epsilon > 0$ such that the interval $(x - \epsilon, x + \epsilon)$ is contained in U. See Figure 4.3.

Remark 4.2 The interval $(x - \epsilon, x + \epsilon)$ is frequently termed a *neighborhood* of x.

EXAMPLE 4.3 The set $U = \{x \in \mathbb{R} : |x - 3| < 2\}$ is open. To see this, choose a point $x \in U$. Let $\epsilon = 2 - |x - 3| > 0$. [Notice that we are choosing ϵ to be the distance of x to the boundary of U.] Then we claim that the interval $I = (x - \epsilon, x + \epsilon) \subseteq U$.

Figure 4.1: A coffee cup and a donut.
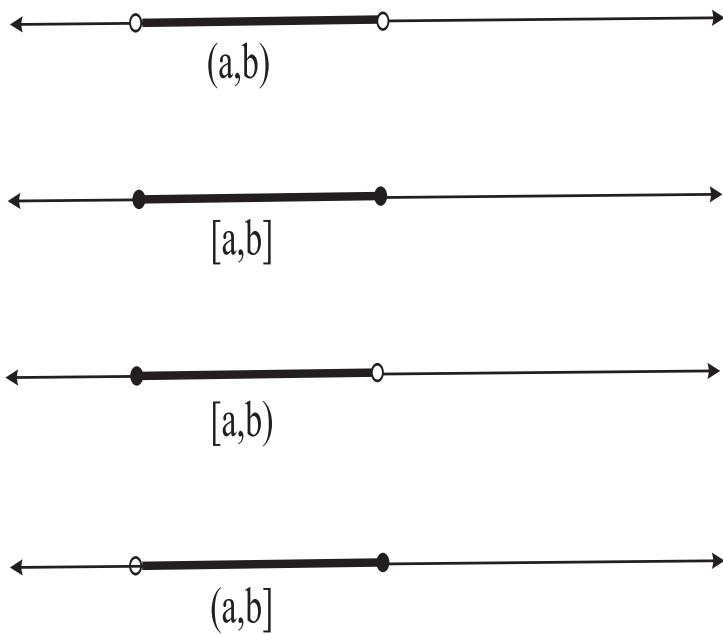
Figure 4.2: Intervals.
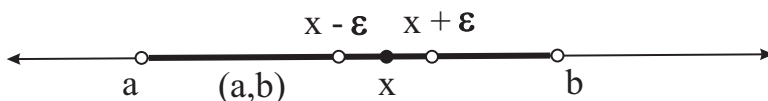
Figure 4.3: An open set.
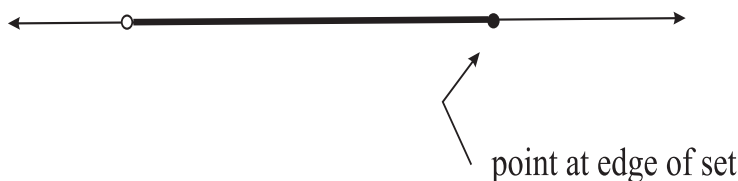
Figure 4.4: A set that is not open.

For, if $t \in I$, then

$$
\begin{aligned}
|t - 3| &\leq |t - x| + |x - 3| \\
&< \epsilon + |x - 3| \\
&= (2 - |x - 3|) + |x - 3| = 2 \,.
\end{aligned}
$$

But this means that $t \in U$.

We have shown that $t \in I$ implies $t \in U$. Therefore $I \subseteq U$. It follows from the definition that U is open. □

Remark 4.4 The way to think about the definition of open set is that a set is open when none of its elements is at the "edge" of the set—each element is surrounded by other elements of the set, indeed a whole interval of them. See Figure 4.3 and contrast it with Figure 4.4. The remainder of this section will make these comments precise.

Proposition 4.5 *If U_α are open sets, for α in some (possibly uncountable) index set A, then*

$$
U = \bigcup_{\alpha \in A} U_\alpha
$$

is open.

Proof: Let $x \in U$. By definition of union, the point x must lie in some U_α. But U_α is open. Therefore there is an interval $I = (x - \epsilon, x + \epsilon)$ such that $I \subseteq U_\alpha$. Therefore $I \subseteq U$. This proves that U is open. □

Figure 4.5: Structure of an open set.

Proposition 4.6 *If $U_1, U_2, \ldots, U_k$ are open sets then the set*

$$V = \bigcap_{j=1}^{k} U_j$$

is also open.

Proof: Let $x \in V$. Then $x \in U_j$ for each j. Since each U_j is open there is for each j a positive number ϵ_j such that $I_j = (x - \epsilon_j, x + \epsilon_j)$ lies in U_j. Set $\epsilon = \min\{\epsilon_1, \ldots, \epsilon_k\}$. Then $\epsilon > 0$ and $(x - \epsilon, x + \epsilon) \subseteq U_j$ for every j. But that just means that $(x - \epsilon, x + \epsilon) \subseteq V$. Therefore V is open. $\square$

Notice the difference between these two propositions: arbitrary unions of open sets are open. But, in order to guarantee that an intersection of open sets is still open, we had to assume that we were only intersecting finitely many such sets. If there were infinitely many sets then the minimum of the ϵ_j could be 0.

To understand this matter better, bear in mind the example of the open sets

$$U_j = \left(-\frac{1}{j}, \frac{1}{j}\right) \;, \quad j = 1, 2, \ldots .$$

Each of the sets U_j is open, but the intersection of the sets U_j is the singleton $\{0\}$, which is *not* open.

The same analysis as in the first example shows that, if $a < b$, then the interval (a, b) is an open set. On the other hand, intervals of the form $(a, b]$ or $[a, b)$ or $[a, b]$ are *not* open. In the first instance, the point b is the center of no interval $(b - \epsilon, b + \epsilon)$ contained in $(a, b]$. Think about the other two intervals to understand why they are not open. We call intervals of the form (a, b) *open intervals*.

We are now in a position to give a complete description of all open sets.

Proposition 4.7 *Let $U \subseteq \mathbb{R}$ be a nonempty open set. Then there are either finitely many or countably many pairwise disjoint open intervals I_j such that*

$$U = \bigcup_{j=1}^{\infty} I_j \, .$$

See Figure 4.5.

Proof: Assume that U is an open subset of the real line. We define an equivalence relation on the set U. The resulting equivalence classes will be the open intervals I_j.

Figure 4.6: A closed set.

Let a and b be elements of U. We say that a is related to b if all real numbers between a and b are also elements of U. It is obvious that this relation is both reflexive and symmetric. For transitivity notice that if a is related to b and b is related to c then (assuming that a, b, c are distinct) one of the numbers a, b, c must lie between the other two. Assume for simplicity that $a < b < c$. Then all numbers between a and c lie in U, for all such numbers are either between a and b or between b and c or are b itself. Thus a is related to c. (The other possible orderings of a, b, c are left for you to consider.)

Thus we have an equivalence relation on the set U. Call the equivalence classes $\{U_\alpha\}_{\alpha \in A}$. We claim that each U_α is an open interval. In fact if a, b are elements of some U_α then all points between a and b are in U. But then a moment's thought shows that each of those "in between" points is related to both a and b. Therefore all points between a and b are elements of U_α. We conclude that U_α is an interval. Is it an *open* interval?

Let $x \in U_\alpha$. Then $x \in U$ so that there is an open interval $I = (x - \epsilon, x + \epsilon)$ contained in U. But x is related to all the elements of I; it follows that $I \subseteq U_\alpha$. Therefore U_α is open.

We have exhibited the set U as a union of open intervals. These intervals are pairwise disjoint because they arise as the equivalence classes of an equivalence relation. Finally, each of these open intervals contains a (different) rational number (why?). Therefore there can be at most countably many of the intervals U_α. ☐

It is worth noting, and we shall learn more about this fact in Chapter 12, that there is no structure theorem for open sets (like the one that we just proved) in dimension 2 and higher. The geometry of Euclidean space gets *much* more complicated as the dimension increases.

Definition 4.8 A subset $F \subseteq \mathbb{R}$ is called *closed* if the complement $\mathbb{R} \setminus F$ is open. See Figure 4.6.

EXAMPLE 4.9 The set $[0, 1]$ is closed. For its complement is

$$(-\infty, 0) \cup (1, \infty),$$

which is open.

EXAMPLE 4.10 An interval of the form $[a, b] = \{x : a \le x \le b\}$ is closed. For its complement is $(-\infty, a) \cup (b, \infty)$, which is the union of two open intervals.

The finite set $A = \{-4, -2, 5, 13\}$ is closed because its complement is

$$(-\infty, -4) \cup (-4, -2) \cup (-2, 5) \cup (5, 13) \cup (13, \infty),$$

which is open.

The set $B = \{1, 1/2, 1/3, 1/4, \dots\} \cup \{0\}$ is closed, for its complement is the set

$$(-\infty, 0) \cup \left\{ \bigcup_{j=1}^{\infty} (1/(j+1), 1/j) \right\} \cup (1, \infty),$$

which is open.

Verify for yourself that if the point 0 is omitted from the set B, then the set is no longer closed.   □

Roughly speaking, a closed set is a set that contains all its limit points. An open set is just the opposite—open sets tend not to contain their limit points. The discussion below will make these ideas more precise.

Remark 4.11 A common mistake that students make is to suppose that every set is either open or closed. This is not true. For instance, the set $[0, 1) = \{x \in \mathbb{R} : 0 \leq x < 1\}$ is neither open nor closed.

Proposition 4.12 *If E_α are closed sets, for α in some (possibly uncountable) index set A, then*

$$E = \bigcap_{\alpha \in A} E_\alpha$$

is closed.

Proof: This is just the contrapositive of Proposition 4.5 above: if U_α is the complement of E_α, each α, then U_α is open. Then $U = \cup U_\alpha$ is also open. But then

$$E = \bigcap E_\alpha = \bigcap {}^c (U_\alpha) = {}^c \left(\bigcup U_\alpha \right) = {}^c U$$

is closed. Here ${}^c S$ denotes the complement of a set S.   □

The fact that the set B in the last example is closed, but that $B \setminus \{0\}$ is not, is placed in perspective by the next proposition.

Proposition 4.13 *Let S be a set of real numbers. Then S is closed if and only if every Cauchy sequence $\{s_j\}$ of elements of S has a limit point which is also an element of S.*

Proof: First suppose that S is closed and let $\{s_j\}$ be a Cauchy sequence in S. We know, since the reals are complete, that there is an element $s \in \mathbb{R}$ such that $s_j \to s$. The point of this half of the proof is to see that $s \in S$. If this statement were false then $s \in T = \mathbb{R} \setminus S$. But T must be open since it is the complement of a closed set. Thus there is an $\epsilon > 0$ such that the interval $I = (s - \epsilon, s + \epsilon) \subseteq T$. This means that no element of S lies in I. In particular, $|s - s_j| \geq \epsilon$ for every j. This contradicts the statement that $s_j \to s$. We conclude that $s \in S$.

Conversely, assume that every Cauchy sequence in S has its limit in S. If S were not closed then its complement would not be open. Hence there would
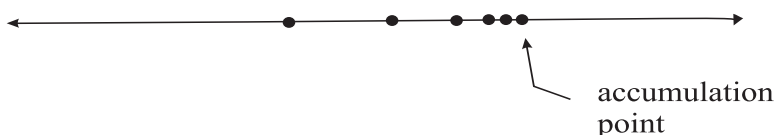
Figure 4.7: The idea of an accumulation point.

be a point $t \in \mathbb{R} \setminus S$ with the property that no interval $(t - \epsilon, t + \epsilon)$ lies in $\mathbb{R} \setminus S$. In other words, $(t - \epsilon, t + \epsilon) \cap S \neq \emptyset$ for every $\epsilon > 0$. Thus for $j = 1, 2, 3, \ldots$ we may choose a point $s_j \in (t - 1/j, t + 1/j) \cap S$. It follows that $\{s_j\}$ is a sequence of elements of S that converge to $t \in \mathbb{R} \setminus S$. That contradicts our hypothesis. We conclude that S must be closed. □

Definition 4.14 Let S be a subset of $\mathbb{R}$. A point x is called an *accumulation point* of S if every neighborhood of x contains infinitely many distinct elements of S. See Figure 4.7. In particular, x is an accumulation point of S if it is the limit of a sequence of distinct elements in S.

The last proposition tells us that closed sets are characterized by the property that they contain all of their accumulation points.

Exercises

1. Let S be any set and $\epsilon > 0$. Define $T = \{t \in \mathbb{R} : |t - s| < \epsilon \text{ for some } s \in S\}$. Prove that T is open.

2. Let S be any set and define $V = \{t \in \mathbb{R} : |t - s| \leq 1 \text{ for some } s \in S\}$. Is V necessarily closed?

3. Let S be a set of real numbers. If S is not open then must it be closed? If S is not closed then must it be open?

4. The *closure* of a set S is the intersection of all closed sets that contain S. Call a set S *robust* if it is the closure of its interior. Which sets of reals are robust?

5. Give an example of nonempty *closed* sets $X_1 \supseteq X_2 \supseteq \ldots$ such that $\cap_j X_j = \emptyset$.

6. Give an example of nonempty closed sets $X_1 \subseteq X_2 \ldots$ such that $\cup_j X_j$ is open.

7. Give an example of open sets $U_1 \supseteq U_2 \ldots$ such that $\cap_j U_j$ is closed and nonempty.

8. Exhibit a countable collection of open sets U_j such that each open set $\mathcal{O} \subseteq \mathbb{R}$ can be written as a union of some of the sets U_j.

a boundary point

Figure 4.8: The idea of a boundary point.

9. Let $S \subseteq \mathbb{R}$ be the rational numbers. Is S open? Is S closed?

10. Let S be an uncountable subset of $\mathbb{R}$. Prove that S must have infinitely many accumulation points. Must it have uncountably many?

* **11.** Let S be any set and define, for $x \in \mathbb{R}$,

$$\text{dis}(x, S) = \inf\{|x - s| : s \in S\}\,.$$

Prove that, if $x \notin \overline{S}$, then $\text{dis}(x, S) > 0$. If $x, y \in \mathbb{R}$ then prove that

$$|\text{dis}(x, S) - \text{dis}(y, S)| \leq |x - y|\,.$$

4.2   Further Properties of Open and Closed Sets

Definition 4.15 Let $S \subseteq \mathbb{R}$ be a set. We call $b \in \mathbb{R}$ a *boundary point* of S if every nonempty neighborhood $(b - \epsilon, b + \epsilon)$ contains both points of S and points of $\mathbb{R} \setminus S$. See Figure 4.8. We denote the set of boundary points of S by ∂S.

A boundary point b might lie in S and might lie in the complement of S. The next example serves to illustrate the concept:

EXAMPLE 4.16 Let S be the interval $(0, 1)$. Then no point of $(0, 1)$ is in the boundary of S since every point of $(0, 1)$ has a neighborhood that lies entirely inside $(0, 1)$. Also, no point of the complement of $T = [0, 1]$ lies in the boundary of S for a similar reason. Indeed, the only candidates for elements of the boundary of S are 0 and 1. See Figure 4.9. The point 0 *is* an element of the boundary since every neighborhood $(0 - \epsilon, 0 + \epsilon)$ contains the point $\epsilon/2 \in S$ and the point $-\epsilon/2 \in \mathbb{R} \setminus S$. A similar calculation shows that 1 lies in the boundary of S.

Now consider the set $T = [0, 1]$. Certainly there are no boundary points in $(0, 1)$, for the same reason as in the first paragraph. And there are no boundary points in $\mathbb{R} \setminus [0, 1]$, since that set is open. Thus the only candidates for elements of the boundary are 0 and 1. As in the first paragraph, these are both indeed boundary points for T. See Figure 4.10.

Notice that neither of the boundary points of S lie in S while both of the boundary points of T lie in T. $\square$

The collection of all boundary points of a set S is called the *boundary* of S and is denoted by ∂S.
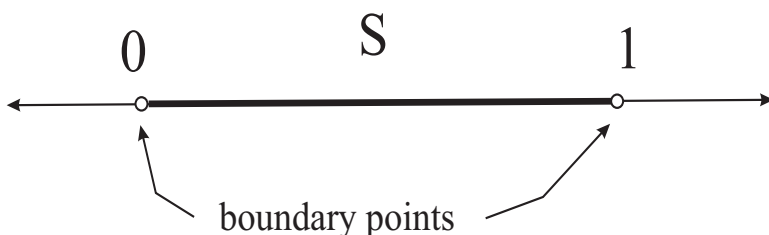
Figure 4.9: Boundary of the open unit interval.
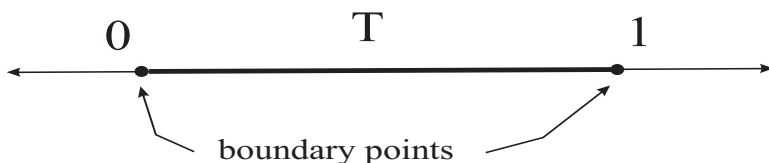
Figure 4.10: Boundary of the closed unit interval.

EXAMPLE 4.17 The boundary of the set $\mathbb{Q}$ is the entire real line. For if x is any element of $\mathbb{R}$ then every interval $(x - \epsilon, x + \epsilon)$ contains both rational numbers and irrational numbers. □

The union of a set S with its boundary is called the *closure* of S, denoted $\overline{S}$. The next example illustrates the concept.

EXAMPLE 4.18 Let S be the set of rational numbers in the interval $[0, 1]$. Then the closure $\overline{S}$ of S is the entire interval $[0, 1]$.

Let T be the open interval $(0, 1)$. Then the closure $\overline{T}$ of T is the closed interval $[0, 1]$. □

Definition 4.19 Let $S \subseteq \mathbb{R}$. A point $s \in S$ is called an *interior point* of S if there is an $\epsilon > 0$ such that the interval $(s - \epsilon, s + \epsilon)$ lies in S. See Figure 4.11. We call the set of all interior points the *interior* of S, and we denote this set by $\overset{\circ}{S}$.

Definition 4.20 A point $t \in S$ is called an *isolated point* of S if there is an $\epsilon > 0$ such that the intersection of the interval $(t - \epsilon, t + \epsilon)$ with S is just the singleton $\{t\}$. See Figure 4.12.

By the definitions given here, an isolated point t of a set $S \subseteq \mathbb{R}$ is a boundary point. For any interval $(t - \epsilon, t + \epsilon)$ contains a point of S (namely, t itself) and points of $\mathbb{R} \setminus S$ (since t is isolated).

Proposition 4.21 *Let $S \subseteq \mathbb{R}$. Then each point of S is either an interior point or a boundary point of S.*

0

an interior point of the interval
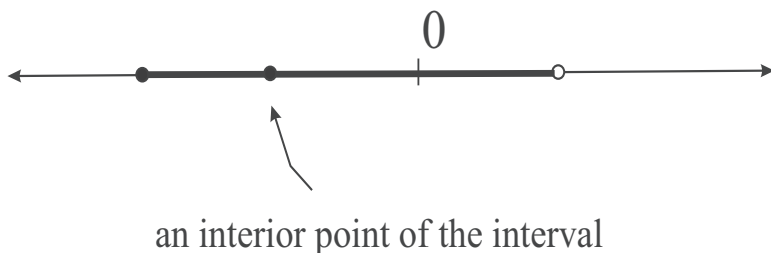
Figure 4.11: The idea of an interior point.

isolated points

Figure 4.12: The idea of an isolated point.

Proof: Fix $s \in S$. If s is not an interior point then no open interval centered at s contains only elements of s. Thus any interval centered at s contains an element of S (namely, s itself) and also contains points of $\mathbb{R} \setminus S$. Thus s is a boundary point of S. □

EXAMPLE 4.22 Let $S = [0, 1]$. Then the interior points of S are the elements of $(0, 1)$. The boundary points of S are the points 0 and 1. The set S has no isolated points.

Let $T = \{1, 1/2, 1/3, \ldots\} \cup \{0\}$. Then the points $1, 1/2, 1/3, \ldots$ are isolated points of T. The point 0 is an accumulation point of T. Every element of T is a boundary point, and there are no others. □

Remark 4.23 Observe that the interior points of a set S are *elements* of S—by their very definition. Also isolated points of S are elements of S. However, a boundary point of S may or may not be an element of S.

If x is an accumulation point of S then every open neighborhood of x contains infinitely many elements of S. Hence x is either a boundary point of S or an interior point of S; it *cannot* be an isolated point of S.

Proposition 4.24 *Let S be a subset of the real numbers. Then the boundary of S equals the boundary of $\mathbb{R} \setminus S$.*

Proof: If x is in the boundary of S, then any neighborhood of x contains points of S and points of ^cS. Thus every neighborhood of x contains points of ^cS and points of S. So x is in the boundary of ^cS. □

The next theorem allows us to use the concept of boundary to distinguish open sets from closed sets.

Figure 4.13: A bounded set.

Theorem 4.25 *A closed set contains all of its boundary points. An open set contains none of its boundary points.*

Proof: Let S be closed and let x be an element of its boundary. If every neighborhood of x contains points of S *other than x itself* then x is an accumulation point of S hence $x \in S$. If not every neighborhood of x contains points of S other than x itself, then there is an $\epsilon > 0$ such that $\{(x-\epsilon, x) \cup (x, x+\epsilon)\} \cap S = \emptyset$. The only way that x can be an element of ∂S in this circumstance is if $x \in S$. That is what we wished to prove.

For the other half of the theorem notice that if T is open then ${}^c T$ is closed. But then ${}^c T$ will contain all its boundary points, which are the same as the boundary points of T itself (why is this true?). Thus T can contain none of its boundary points. □

Proposition 4.26 *Every nonisolated boundary point of a set S is an accumulation point of the set S.*

Proof: This proof is treated in the exercises. □

Definition 4.27 A subset S of the real numbers is called *bounded* if there is a positive number M such that $|s| \leq M$ for every element s of S. See Figure 4.13.

The next result is one of the great theorems of nineteenth century analysis. It is essentially a restatement of the Bolzano–Weierstrass theorem of Section 2.2.

Theorem 4.28 (Bolzano–Weierstrass) *Every bounded, infinite subset of $\mathbb{R}$ has an accumulation point.*

Proof: Let S be a bounded, infinite set of real numbers. Let $\{a_j\}$ be a sequence of distinct elements of S. By Theorem 2.24, there is a subsequence $\{a_{j_k}\}$ that converges to a limit α. Then α is an accumulation point of S. □

Corollary 4.29 *Let $S \subseteq \mathbb{R}$ be a nonempty, closed, and bounded set. If $\{a_j\}$ is any sequence in S, then there is a Cauchy subsequence $\{a_{j_k}\}$ that converges to an element of S.*

Proof: Merely combine the Bolzano–Weierstrass theorem with Proposition 4.13 of the last section. □

EXAMPLE 4.30 Consider the set $\{\sin j\}$. This set of real numbers is bounded by 1. By the Bolzano–Weierstrass theorem, it therefore has an accumulation point. So there is a sequence $\{\sin j_k\}$ that converges to some limit point p, even though it would be difficult to say precisely what that sequence is. □

Exercises

1. Let S be any set of real numbers. Prove that $S \subseteq \overline{S}$. Prove that $\overline{S}$ is a closed set. Prove that $\overline{S} \setminus \overset{\circ}{S}$ is the boundary of S.

2. What is the interior of the set $S = \{1, 1/2, 1/3, \dots\} \cup \{0\}$? What is the boundary of the set?

3. The union of infinitely many closed sets need not be closed. It need not be open either. Give examples to illustrate the possibilities.

4. The intersection of infinitely many open sets need not be open. It need not be closed either. Give examples to illustrate the possibilities.

5. Let S be any set of real numbers. Prove that $\overset{\circ}{S}$ is open. Prove that S is open if and only if S equals its interior.

6. Prove Proposition 4.26.

7. Let $S \subseteq \mathbb{R}$ be the rational numbers. What is the interior of S? What is the boundary of S? What is the closure of S?

* 8. Give an example of a one-to-one, onto, continuous function f with a continuous inverse from the halfline $(0, \infty)$ to the full line $(-\infty, \infty)$.

* 9. Give an example of a closed set in the plane (refer to Chapter 12) whose projection on the x-axis is not closed.

* 10. Show that the projection of an open set in the plane (refer to Chapter 12) into the x-axis must be open.

4.3 Compact Sets

Compact sets are sets (usually infinite) which share many of the most important properties of finite sets. They play an important role in real analysis.

Definition 4.31 A set $S \subseteq \mathbb{R}$ is called *compact* if every sequence in S has a subsequence that converges *to an element of S*.

Theorem 4.32 (Heine–Borel) *A set $S \subseteq \mathbb{R}$ is compact if and only if it is closed and bounded.*

an open cover of the set S
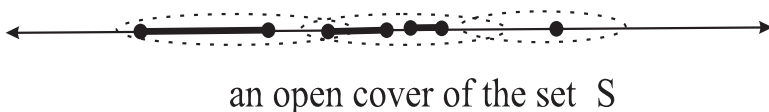
Figure 4.14: Open covers and compactness.

Proof: That a closed, bounded set has the property of compactness is the content of Corollary 4.29 and Proposition 4.13.

Now let S be a set that is compact. If S is not bounded, then there is an element s_1 of S that has absolute value larger than 1. Also there must be an element s_2 of S that has absolute value larger than 2. Continuing, we find elements $s_j \in S$ satisfying

$$|s_j| > j$$

for each j. But then no subsequence of the sequence $\{s_j\}$ can be Cauchy. This contradiction shows that S must be bounded.

If S is compact but S is not closed, then there is a point x which is the limit of a sequence $\{s_j\} \subseteq S$ but which is not itself in S. But every sequence in S is, by definition of "compact," supposed to have a subsequence converging *to an element of S*. For the sequence $\{s_j\}$ that we are considering, x is the only candidate for the limit of a subsequence. Thus it must be that $x \in S$. That contradiction establishes that S is closed. $\qquad\square$

In the abstract theory of topology (where there is no notion of distance), sequences cannot be used to characterize topological properties. Therefore a different definition of compactness is used. For interest's sake, and for future use, we now show that the definition of compactness that we have been discussing is equivalent to the one used in abstract topology theory. First we need a new definition.

Definition 4.33 Let S be a subset of the real numbers. A collection of open sets $\{\mathcal{O}_\alpha\}_{\alpha \in A}$ (each $\mathcal{O}_\alpha$ is an open set of real numbers) is called an *open covering* of S if

$$\bigcup_{\alpha \in A} \mathcal{O}_\alpha \supseteq S.$$

See Figure 4.14.

EXAMPLE 4.34 The collection $\mathcal{C} = \{(1/j, 1)\}_{j=1}^{\infty}$ is an open covering of the interval $I = (0, 1)$. No finite subcollection of the elements of $\mathcal{C}$ covers I.

The collection $\mathcal{D} = \{(1/j, 1)\}_{j=1}^{\infty} \cup \{(-1/5, 1/5), (4/5, 6/5)\}$ is an open covering of the interval $J = [0, 1]$. However, not all the elements $\mathcal{D}$ are actually needed to cover J. In fact

$$(-1/5, 1/5) \,, \ (1/6, 1) \,, \ (4/5, 6/5)$$

cover the interval J. $\qquad\square$

It is the distinction displayed in this example that distinguishes compact sets from the point of view of topology. To understand the point, we need another definition:

Definition 4.35 If $\mathcal{C}$ is an open covering of a set S and if $\mathcal{D}$ is another open covering of S such that each element of $\mathcal{D}$ is also an element of $\mathcal{C}$ then we call $\mathcal{D}$ a *subcovering* of $\mathcal{C}$.

We call $\mathcal{D}$ a *finite subcovering* if $\mathcal{D}$ has just finitely many elements.

EXAMPLE 4.36 The collection of intervals

$$\mathcal{C} = \{(j-1, j+1)\}_{j=1}^{\infty}$$

is an open covering of the set $S = [5, 9]$. The collection

$$\mathcal{D} = \{(j-1, j+1)\}_{j=5}^{\infty}$$

is a subcovering.

However, the collection

$$\mathcal{E} = \{(4, 6), (5, 7), (6, 8), (7, 9), (8, 10)\}$$

is a *finite* subcovering.   □

Theorem 4.37 *A set $S \subseteq \mathbb{R}$ is compact if and only if every open covering $\mathcal{C} = \{\mathcal{O}_\alpha\}_{\alpha \in A}$ of S has a finite subcovering.*

Proof: Assume that S is a compact set and let $\mathcal{C} = \{\mathcal{O}_\alpha\}_{\alpha \in A}$ be an open covering of S.

By Theorem 4.29, S is closed and bounded. Therefore it holds that $a = \inf S$ is a finite real number, and an element of S. Likewise, $b = \sup S$ is a finite real number and an element of S. Write $I = [a, b]$. The case $a = b$ is trivial so we assume that $a < b$.

Set

$$\mathcal{A} = \{x \in I : \mathcal{C} \text{ contains a finite subcover that covers } S \cap [a, x]\}.$$

Then $\mathcal{A}$ is nonempty since $a \in \mathcal{A}$. Let $t = \sup \mathcal{A}$. Then some element $\mathcal{O}_0$ of $\mathcal{C}$ contains t. Let s be an element of $\mathcal{O}_0$ to the left of t. Then, by the definition of t, s is an element of $\mathcal{A}$. So there is a finite subcovering $\mathcal{C}'$ of $\mathcal{C}$ that covers $S \cap [a, s]$. But then $\mathcal{D} = \mathcal{C}' \cup \{\mathcal{O}_0\}$ covers $S \cap [a, t]$, showing that $t = \sup \mathcal{A}$ lies in A. But in fact $\mathcal{D}$ even covers points to the right of t. Thus t cannot be the supremum of $\mathcal{A}$ unless $t = b$.

We have learned that t must be the point b itself and that therefore $b \in \mathcal{A}$. But that says that $S \cap [a, b] = S$ can be covered by finitely many of the elements of $\mathcal{C}$. That is what we wished to prove.

For the converse, assume that every open covering of S has a finite subcovering. Let $\{a_j\}$ be a sequence in S. Assume, seeking a contradiction, that the sequence has no subsequence that converges to an element of S. This must mean

that for every $s \in S$ there is an $\epsilon_s > 0$ such that no element of the sequence satisfies $0 < |a_j - s| < \epsilon_s$. Let $I_s = (s - \epsilon_s, s + \epsilon_s)$. The collection $\mathcal{C} = \{I_s\}$ is then an open covering of the set S. By hypothesis, there exists a finite subcovering $I_{s_1}, \dots I_{s_k}$ of open intervals that cover S. But then $S \subseteq \cup_{j=1}^k I_{s_j}$ contains no element of the sequence $\{a_j\}$, and that is a contradiction. $\qquad\square$

EXAMPLE 4.38 If $A \subseteq B$ and both sets are nonempty then $A \cap B = A \neq \emptyset$. A similar assertion holds when intersecting *finitely many* nonempty sets $A_1 \supseteq A_2 \supseteq \cdots \supseteq A_k$; it holds in this circumstance that $\cap_{j=1}^k A_j = A_k$.

However, it is possible to have infinitely many nonempty nested sets with null intersection. An example is the sets $I_j = (0, 1/j)$. We see that $I_1 \supseteq I_2 \supseteq I_3 \supseteq \cdots$ yet

$$\bigcap_{j=1}^\infty I_j = \emptyset \,.$$

By contrast, if we take $K_j = [0, 1/j]$ then

$$\bigcap_{j=1}^\infty K_j = \{0\} \,.$$

The next proposition shows that compact sets have the intuitively appealing property of the K_js rather than the unsettling property of the I_js. $\qquad\square$

Proposition 4.39 *Let*

$$K_1 \supseteq K_2 \supseteq \cdots \supseteq K_j \supseteq \dots$$

be nonempty compact sets of real numbers. Set

$$\mathcal{K} = \bigcap_{j=1}^\infty K_j \,.$$

Then $\mathcal{K}$ is compact and $\mathcal{K} \neq \emptyset$.

Proof: Each K_j is closed and bounded hence $\mathcal{K}$ is closed and bounded. Thus $\mathcal{K}$ is compact. Let $x_j \in K_j$, each j. Then $\{x_j\} \subseteq K_1$. By compactness, there is a convergent subsequence $\{x_{j_k}\}$ with limit $x_0 \in K_1$. However, $\{x_{j_k}\}_{k=2}^\infty \subseteq K_2$. Thus $x_0 \in K_2$. Similar reasoning shows that $x_0 \in K_m$ for all $m = 1, 2, \dots$. In conclusion, $x_0 \in \cap_j K_j = \mathcal{K}$. $\qquad\square$

Exercises

   1. Prove that the intersection of a compact set and a closed set is compact.

2. Let K be a compact set and let U be an open set that contains K. Prove that there is an $\epsilon > 0$ such that, if $k \in K$, then the interval $(k - \epsilon, k + \epsilon)$ is contained in U.

3. Let K be compact and L closed, and assume that the two sets are disjoint. Show that there is a positive distance between the two sets.

4. Let K be a compact set. Let $\delta > 0$. Prove that there is a finite collection of intervals of radius δ that covers K.

5. Let K be a compact set. Let $\mathcal{U} = \{U_j\}_{j=1}^{k}$ be a finite open covering of K. Show that there is a $\delta > 0$ so that, if x is any point of K, then the disc or interval of center x and radius δ lies entirely in one of the U_j.

6. Assume that we have intervals $[a_1, b_1] \supseteq [a_2, b_2] \supseteq \cdots$, each of positive length, and that $\lim_{j \to \infty} |a_j - b_j| = 0$. Prove that there is a point x such that $x \in [a_j, b_j]$ for every j.

7. Let $K \subseteq \mathbb{R}$ be a compact set. Let $\epsilon > 0$. Define

$$\widehat{K} = \{t \in \mathbb{R} : |t - k| \leq \epsilon \text{ for some } k \in K\}.$$

   Is $\widehat{K}$ compact? Why or why not?

8. If K in $\mathbb{R}$ is compact then show that $^c K$ is not compact.

9. Prove that the intersection of any number of compact sets is compact. The analogous statement for unions is false.

10. Let $U \subset \mathbb{R}$ be any open set. Show that there exist compact sets $K_1 \subset K_2 \subset \cdots$ so that $\cup_j K_j = U$.

11. Produce an open set U in the real line so that U may not be written as the decreasing intersection of compact sets.

4.4   The Cantor Set

In this section we describe the construction of a remarkable subset of $\mathbb{R}$ with many pathological properties. It only begins to suggest the richness of the structure of the real number system.

We begin with the unit interval $S_0 = [0, 1]$. We extract from S_0 its open middle third; thus $S_1 = S_0 \setminus (1/3, 2/3)$. Observe that S_1 consists of two closed intervals of equal length $1/3$. See Figure 4.15.

Now we construct S_2 from S_1 by extracting from each of its two intervals the middle third: $S_2 = [0, 1/9] \cup [2/9, 1/3] \cup [6/9, 7/9] \cup [8/9, 1]$. Figure 4.16 shows S_2.
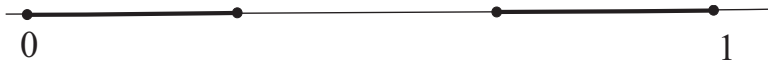
Figure 4.15: Construction of the Cantor set.
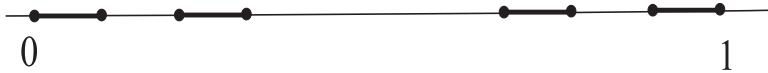
Figure 4.16: Second step in the construction of the Cantor set.

Continuing in this fashion, we construct S_{j+1} from S_j by extracting the middle third from each of its component subintervals. We define the Cantor set C to be

$$C = \bigcap_{j=1}^{\infty} S_j \, .$$

Notice that each of the sets S_j is closed and bounded, hence compact. By Proposition 4.39 of the last section, C is therefore not empty (one can also note that the endpoints of the removed intervals are all in the Cantor set). The set C is closed and bounded, hence compact.

Proposition 4.40 *The Cantor set C has zero length, in the sense that the complementary set $[0,1] \setminus C$ has length 1.*

Proof: In the construction of S_1, we removed from the unit interval one interval of length 3^{-1}. In constructing S_2, we further removed two intervals of length 3^{-2}. In constructing S_j, we removed 2^{j-1} intervals of length 3^{-j}. Thus the total length of the intervals removed from the unit interval is

$$\sum_{j=1}^{\infty} 2^{j-1} \cdot 3^{-j} \, .$$

This last equals

$$\frac{1}{3} \sum_{j=0}^{\infty} \left(\frac{2}{3} \right)^j \, .$$

The geometric series sums easily and we find that the total length of the intervals removed is

$$\frac{1}{3} \left(\frac{1}{1 - 2/3} \right) = 1 \, .$$

Thus the Cantor set has length zero because its complement in the unit interval has length one.   □

Proposition 4.41 *The Cantor set is uncountable.*

first digit 0   **first digit 1**

0   1

Figure 4.17: The first digit of the label of a point in the Cantor set.

second digit 0   second digit 1   second digit 0   second digit 1

0   1
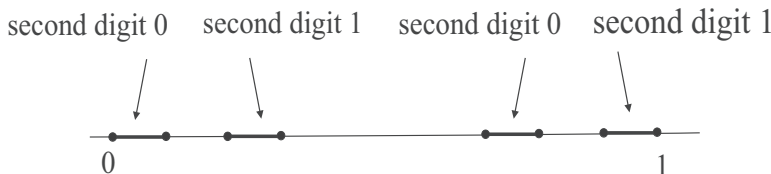
Figure 4.18: The second digit of the label of a point in the Cantor set.

Proof: We assign to each element of the Cantor set a "label" consisting of a sequence of 0s and 1s that identifies its location in the set.

Fix an element x in the Cantor set. Then x is in S_1. If x is in the left half of S_1, then the first digit in the "label" of x is 0; otherwise it is 1. See Figure 4.17.

Likewise $x \in S_2$. By the first part of this argument, it is either in the left half S_2^0 of S_2 (when the first digit in the label is 0) or the right half S_2^1 of S_2 (when the first digit of the label is 1). Whichever of these is correct, that half will consist of two intervals of length 3^{-2}. If x is in the leftmost of these two intervals then the second digit of the "label" of x is 0. Otherwise the second digit is 1. Refer to Figure 4.18.

Continuing in this fashion, we may assign to x an infinite sequence of 0s and 1s.

Conversely, if $a, b, c, \ldots$ is a sequence of 0s and 1s, then we may locate a unique corresponding element y of the Cantor set. If the first digit is a zero then y is in the left half of S_1; otherwise y is in the right half of S_1. Likewise the second digit locates y within S_2, and so forth.

Thus we have a one-to-one correspondence between the Cantor set and the collection of all infinite sequences of zeroes and ones. [Notice that we are in effect thinking of the point assigned to a sequence $c_1 c_2 c_3 \ldots$ of 0s and 1s as the limit of the points assigned to $c_1, c_1 c_2, c_1 c_2 c_3, \ldots$ Thus we are using the fact that C is closed.] However, as we can learn in Appendix I at the end of the book, the set of all infinite sequences of zeroes and ones is uncountable. Thus we see that the Cantor set is uncountable.   □

The Cantor set is quite thin (it has zero length) but it is large in the sense that it has uncountably many elements. Also it is compact. The next result

reveals a surprising, and not generally well known, property of this "thin" set.

Theorem 4.42 *Let C be the Cantor set and define*

$$S = \{x + y : x \in C, y \in C\}.$$

Then $S = [0, 2]$.

Proof: We sketch the proof here and treat the details in the exercises.

Since $C \subseteq [0, 1]$ it is clear that $S \subseteq [0, 2]$. For the reverse inclusion, fix an element $t \in [0, 2]$. Our job is to find two elements c and d in C such that $c + d = t$.

First observe that $\{x + y : x \in S_1, y \in S_1\} = [0, 2]$. Therefore there exist $x_1 \in S_1$ and $y_1 \in S_1$ such that $x_1 + y_1 = t$.

Similarly, $\{x + y : x \in S_2, y \in S_2\} = [0, 2]$. Therefore there exist $x_2 \in S_2$ and $y_2 \in S_2$ such that $x_2 + y_2 = t$.

Continuing in this fashion we may find for each j numbers x_j and y_j such that $x_j, y_j \in S_j$ and $x_j + y_j = t$. Of course $\{x_j\} \subseteq C$ and $\{y_j\} \subseteq C$ hence there are subsequences $\{x_{j_k}\}$ and $\{y_{j_k}\}$ which converge to real numbers c and d respectively. Since C is compact, we can be sure that $c \in C$ and $d \in C$. But the operation of addition respects limits, thus we may pass to the limit as $k \to \infty$ in the equation

$$x_{j_k} + y_{j_k} = t$$

to obtain

$$c + d = t.$$

Therefore $[0, 2] \subseteq \{x + y : x \in C\}$. This completes the proof. $\square$

In the exercises at the end of the section we shall explore constructions of other Cantor sets, some of which have zero length and some of which have positive length. The Cantor set that we have discussed in detail in the present section is sometimes distinguished with the name "the Cantor ternary set." We shall also consider in the exercises other ways to construct the Cantor ternary set.

Observe that, whereas any open set is the countable or finite disjoint union of open intervals, the existence of the Cantor set shows us that there is no such structure theorem for closed sets. That is to say, we cannot hope to write an arbitrary closed set as the disjoint union of closed intervals. [However, de Morgan's Law shows that an arbitrary closed set can be written as the countable intersection of sets, each of which is the union of disjoint closed intervals.] In fact closed intervals are atypically simple when considered as examples of closed sets.

Exercises

1. Construct a Cantor-like set by removing the middle *fifth* from the unit interval, removing the middle fifth of each of the remaining intervals, and

so on. What is the length of the set that you construct in this fashion? Is it uncountable? Is it different from the Cantor set constructed in the text?

2. Refer to Exercise **1**. Construct a Cantor set by removing, at the jth step, a middle subinterval of length 3^{-2j+1} from each existing interval. The Cantor-like set that results should have positive length. What is that length? Does this Cantor set have the other properties of the Cantor set constructed in the text?

3. Prove that it is not the case that there is a positive distance between two disjoint open sets.

4. Let $0 < \lambda < 1$. Imitate the construction of the Cantor set to produce a subset of the unit interval whose complement has length λ.

5. How many endpoints of the intervals in the S_j are there in the Cantor set? How many non-endpoints?

6. What is the interior of the Cantor set? What is the boundary?

7. Fix the sequence $a_j = 3^{-j}, j = 1, 2, \ldots$. Consider the set S of all sums

$$\sum_{j=1}^{\infty} \mu_j a_j \,,$$

where each μ_j is one of the numbers 0 or 2. Show that S is the Cantor set. If s is an element of $S, s = \sum \mu_j a_j$, and if $\mu_j = 0$ for all j sufficiently large, then show that s is an endpoint of one of the intervals in one of the sets S_j that were used to construct the Cantor set in the text.

8. Let us examine the proof that $\{x + y : x \in C, y \in C\}$ equals $[0, 2]$ more carefully.

   a) Prove for each j that $\{x+y : x \in S_j, y \in S_j\}$ equals the interval $[0, 2]$.

   b) For $t \in [0, 1]$, explain how the subsequences $\{x_{j_k}\}$ and $\{y_{j_k}\}$ in S_j can be chosen to satisfy $x_{j_k} + y_{j_k} = t$ and so that $x_{j_k} \to x_0 \in C$ and $y_{j_k} \to y_0 \in C$. Observe that it is important for the proof that the index j_k be the same for both subsequences.

   c) Formulate a suitable statement concerning the assertion that the binary operation of addition "respects limits" as required in the argument in the text. Prove this statement and explain how it allows us to pass to the limit in the equation $x_{j_k} + y_{j_k} = t$.

9. Use the characterization of the Cantor set from Exercise **8** to give a new proof of the fact that $\{x + y : x \in C, y \in C\}$ equals the interval $[0, 2]$.
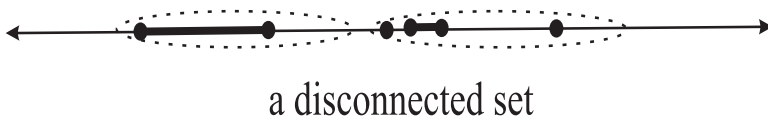
a disconnected set

Figure 4.19: The idea of disconnected.

10. How many points in the Cantor set have finite ternary expansions? How many have infinite ternary expansions?

* 11. Discuss which sequences a_j of positive numbers could be used as in Exercise **8** to construct sets which are like the Cantor set.

* 12. Describe how to produce a two-dimensional Cantor-like set in the plane.

4.5 Connected and Disconnected Sets

Definition 4.43 Let S be a set of real numbers. We say that S is *disconnected* if it is possible to find a pair of open sets U and V such that

$$U \cap S \neq \emptyset, V \cap S \neq \emptyset,$$

$$(U \cap S) \cap (V \cap S) = \emptyset,$$

and

$$S = (U \cap S) \cup (V \cap S).$$

See Figure 4.19. If no such U and V exist then we call S *connected*.

EXAMPLE 4.44 The set $T = \{x \in \mathbb{R} : |x| < 1, x \neq 0\}$ is disconnected. Take $U = \{x : x < 0\}$ and $V = \{x : x > 0\}$. Then

$$U \cap T = \{x : -1 < x < 0\} \neq \emptyset$$

and

$$V \cap T = \{x : 0 < x < 1\} \neq \emptyset.$$

Also $(U \cap T) \cap (V \cap T) = \emptyset$. Clearly $T = (U \cap T) \cup (V \cap T)$, hence T is disconnected.
□

It is clear that a disconnected set has the property that there are disjoint open sets that, in effect, *disconnect* the set. The next example looks at the contrapositive.

EXAMPLE 4.45 The set $X = [-1, 1]$ is connected. To see this, suppose to the contrary that there exist open sets U and V such that $U \cap X \neq \emptyset, V \cap X \neq \emptyset, (U \cap X) \cap (V \cap X) = \emptyset$, and

$$X = (U \cap X) \cup (V \cap X).$$

Figure 4.20: A closed interval is connected.

Choose $a \in U \cap X$ and $b \in V \cap X$. Set

$$\alpha = \sup \left(U \cap [a, b] \right) .$$

Now $[a, b] \subseteq X$ hence $U \cap [a, b]$ is disjoint from V. Thus $\alpha \leq b$. But $^c V$ is closed hence $\alpha \notin V$. It follows that $\alpha < b$.

If $\alpha \in U$ then, because U is open, there exists an $\widetilde{\alpha} \in U$ such that $\alpha < \widetilde{\alpha} < b$. This would mean that we chose α incorrectly. Hence $\alpha \notin U$. But $\alpha \notin U$ and $\alpha \notin V$ means $\alpha \notin X$. On the other hand, α is the supremum of a subset of X (since $a \in X, b \in X$, and X is an interval). Since X is a closed interval, we conclude that $\alpha \in X$. This contradiction shows that X must be connected. $\square$

With small modifications, the discussion in the last example demonstrates that any closed interval is connected (Exercise **1**). See Figure 4.20. Also (see Exercise **2**), we may similarly see that any open interval or half-open interval is connected. In fact the converse is true as well:

Theorem 4.46 *A subset S of $\mathbb{R}$ is connected if and only if S is an interval.*

Proof: If S is not an interval then there exist $a \in S, b \in S$ and a point t between a and b such that $t \notin S$. Define $U = \{x \in \mathbb{R} : x < t\}$ and $V = \{x \in \mathbb{R} : t < x\}$. Then U and V are open and disjoint, $U \cap S \neq \emptyset$, $V \cap S \neq \emptyset$, and

$$S = (U \cap S) \cup (V \cap S) .$$

Thus S is disconnected.

If S is an interval then we prove that it is connected using the methodology of Example 4.45. $\square$

The Cantor set is not connected; indeed it is disconnected in a special sense. Call a set S *totally disconnected* if, for each distinct $x \in S$, $y \in S$, there exist disjoint open sets U and V such that $x \in U, y \in V$, and $S = (U \cap S) \cup (V \cap S)$.

Proposition 4.47 *The Cantor set is totally disconnected.*

Proof: Let $x, y \in C$ be distinct and assume that $x < y$. Set $\delta = |x - y|$. Choose j so large that $3^{-j} < \delta$. Then $x, y \in S_j$, but x and y cannot both be in the same interval of S_j (since the intervals will have length equal to 3^{-j}). It follows that there is a point t between x and y that is not an element of S_j, hence not an element of C. Set $U = \{s : s < t\}$ and $V = \{s : s > t\}$. Then $x \in U \cap C$ hence $U \cap C \neq \emptyset$; likewise $V \cap C \neq \emptyset$. Also $(U \cap C) \cap (V \cap C) = \emptyset$. Finally $C = (C \cap U) \cup (C \cap V)$. Thus C is totally disconnected. $\square$

Exercises

1. Imitate Example 4.45 in the text to prove that any closed interval is connected.

2. Imitate Example 4.45 in the text to prove that any open interval or half-open interval is connected.

3. Give an example of a totally disconnected set $S \subseteq [0, 1]$ such that $\overline{S} = [0, 1]$.

4. Write the real line as the union of two totally disconnected sets.

5. Let $S \subseteq \mathbb{R}$ be a set. Let $s, t \in S$. We say that s and t are in the same *connected component* of S if the entire interval $[s, t]$ lies in S. What are the connected components of the Cantor set? Is it possible to have a set S with countably many connected components? With uncountably many connected components?

6. If A is connected and B is connected then will $A \cap B$ be connected?

7. If A is connected and B is connected then will $A \cup B$ be connected?

8. What can you say about the sum of two disconnected sets? Give some examples.

9. If A is connected and B is disconnected then what can you say about $A \cap B$?

10. If sets U_j form the basis of a topology on a space X (that is to say, each open set in X can be written as a union of some of the U_j) and if each U_j is connected, then what can you say about X?

* 11. If A is connected and B is connected then does it follow that $A \times B$ is connected?

4.6   Perfect Sets

Definition 4.48 A set $S \subseteq \mathbb{R}$ is called *perfect* if it is closed and if every point of S is an accumulation point of S.

The property of being perfect is a rather special one: it means that the set has no isolated points.

EXAMPLE 4.49 Consider the set $S = [0, 2]$. This set is perfect. Because **(i)** it is closed, **(ii)** any interior point is clearly an accumulation point, **(iii)** 0 is the limit of $\{1/j\}$ so is an accumulation point, and **(iv)** 2 is the limit of $\{2 - 1/j\}$ so is an accumulation point.   □

Clearly a closed interval $[a, b]$ is perfect. After all, a point x in the interior of the interval is surrounded by an entire open interval $(x - \epsilon, x + \epsilon)$ of elements of the interval; moreover a is the limit of elements from the right and b is the limit of elements from the left.

EXAMPLE 4.50 The Cantor set, *a totally disconnected set*, is perfect. It is definitely closed. Now fix $x \in C$. Then $x \in S_1$. Thus x is in one of the two intervals composing S_1. One (or perhaps both) of the endpoints of that interval does not equal x. Call that endpoint a_1. Likewise $x \in S_2$. Therefore x lies in one of the intervals of S_2. Choose an endpoint a_2 of that interval which does not equal x. Continuing in this fashion, we construct a sequence $\{a_j\}$. Notice that *each of the elements of this sequence lies in the Cantor set* (why?). Finally, $|x - a_j| \leq 3^{-j}$ for each j. Therefore x is the limit of the sequence. We have thus proved that the Cantor set is perfect.   □

The fundamental theorem about perfect sets tells us that such a set must be rather large. We have

Theorem 4.51 *A nonempty perfect set must be uncountable.*

Proof: Let S be a nonempty perfect set. Since S has accumulation points, it cannot be finite. Therefore it is either countable or uncountable.

Seeking a contradiction, we suppose that S is countable. Write $S = \{s_1, s_2, \dots\}$. Set $U_1 = (s_1 - 1, s_1 + 1)$. Then U_1 is a neighborhood of s_1. Now s_1 is a limit point of S so there must be infinitely many elements of S lying in U_1. We select a bounded open interval U_2 such that $\overline{U}_2 \subseteq U_1$, $\overline{U}_2$ does not contain s_1, and U_2 *does* contain some element of S.

Continuing in this fashion, assume that $s_1, \dots, s_j$ have been selected and choose a bounded interval U_{j+1} such that (i) $\overline{U}_{j+1} \subseteq U_j$, (ii) $s_j \notin \overline{U}_{j+1}$, and (iii) U_{j+1} contains some element of S.

Observe that each set $V_j = \overline{U}_j \cap S$ is closed and bounded, hence compact. Also each V_j is nonempty by construction but V_j does not contain s_{j-1}. It follows that $V = \cap_j V_j$ cannot contain s_1 (since V_2 does not), cannot contain s_2 (since V_3 does not), indeed cannot contain any element of S. Hence V, being a subset of S, is empty. But V is the decreasing intersection of nonempty compact sets, hence cannot be empty!

This contradiction shows that S cannot be countable. So it must be uncountable.   □

Corollary 4.52 *If $a < b$ then the closed interval $[a, b]$ is uncountable.*

Proof: The interval $[a, b]$ is perfect.   □

We also have a new way of seeing that the Cantor set is uncountable, since it is perfect:

Corollary 4.53 *The Cantor set is uncountable.*

Exercises

1. Let $U_1 \subseteq U_2 \ldots$ be open sets and assume that each of these sets has bounded, nonempty complement. Is it true that $\cup_j U_j \neq \mathbb{R}$?

2. Let $X_1, X_2, \ldots$ each be perfect sets and suppose that $X_1 \supseteq X_2 \supseteq \ldots$. Set $X = \cap_j X_j$. Is X perfect?

3. Is the product of perfect sets perfect?

4. If $A \cap B$ is perfect, then what may we conclude about A and B?

5. If $A \cup B$ is perfect, then what may we conclude about A and B?

6. Call a set imperfect if its complement is perfect. Which sets are imperfect? Can you specify a connected imperfect set?

7. What can you say about the interior of a perfect set?

8. What can you say about the boundary of a perfect set?

*   9. Let $S_1, S_2, \ldots$ be closed sets and assume that $\cup_j S_j = \mathbb{R}$. Prove that at least one of the sets S_j has nonempty interior. (**Hint:** Use an idea from the proof that perfect sets are uncountable.)

*   10. Let S be a nonempty set of real numbers. A point x is called a *condensation point* of S if every neighborhood of x contains uncountably many points of S. Prove that the set of condensation points of S is closed. Is it necessarily nonempty? Is it nonempty when S is uncountable?

     If T is an uncountable set, then show that the set of its condensation points is perfect.

*   11. Prove that any closed set can be written as the union of a perfect set and a countable set. (**Hint:** Refer to Exercise **10** above.)

Chapter 5

Limits and Continuity of Functions

5.1 Definition and Basic Properties of the Limit of a Function

In this chapter we are going to treat some topics that you have seen before in your calculus class. However, we shall use the deep properties of the real numbers that we have developed in this text to obtain important new insights. Therefore you should *not* think of this chapter as review. Look at the concepts introduced here with the power of your new understanding of analysis.

Definition 5.1 Let $E \subseteq \mathbb{R}$ be a set and let f be a real-valued function with domain E. Fix a point $P \in \mathbb{R}$ that is an accumulation point of E. Let ℓ be a real number. We say that

$$\lim_{E \ni x \to P} f(x) = \ell$$

if, for each $\epsilon > 0$, there is a $\delta > 0$ such that, when $x \in E$ and $0 < |x - P| < \delta$, then

$$|f(x) - \ell| < \epsilon .$$

In words, we say that the limit as x tends to P of f is equal to ℓ.

The definition makes precise the notion that we can force $f(x)$ to be just as close as we please to ℓ by making x sufficiently close to P. Notice that the definition puts the condition $0 < |x - P| < \delta$ on x, so that x is not allowed to take the value P. In other words we do not look at $x = P$, but rather at x *near* to P.

Also observe that we only consider the limit of f at a point P that is not isolated. In the exercises you will be asked to discuss why it would be nonsensical to use the above definition to study the limit at an isolated point.
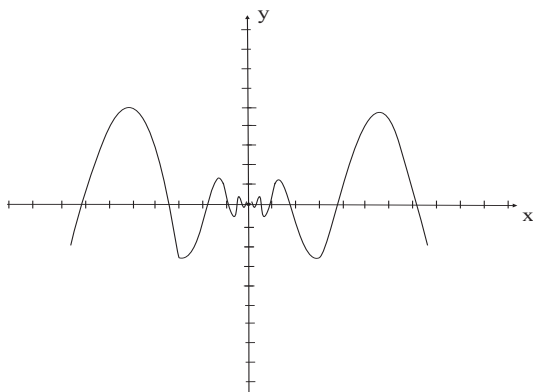
Figure 5.1: The limit of an oscillatory function.

EXAMPLE 5.2 Let $E = \mathbb{R} \setminus \{0\}$ and

$$f(x) = x \cdot \sin(1/x) \ \text{ if } \ x \in E.$$

See Figure 5.1. Then $\lim_{x \to 0} f(x) = 0$. To see this, let $\epsilon > 0$. Choose $\delta = \epsilon$. If $0 < |x - 0| < \delta$ then

$$|f(x) - 0| = |x \cdot \sin(1/x)| \leq |x| < \delta = \epsilon,$$

as desired. Thus the limit exists and equals 0.   □

EXAMPLE 5.3 Let $E = \mathbb{R}$ and

$$g(x) = \begin{cases} 1 & \text{if} \quad x \text{ is rational} \\ 0 & \text{if} \quad x \text{ is irrational.} \end{cases}$$

Then $\lim_{x \to P} g(x)$ does not exist for any point P of E.

To see this, fix $P \in \mathbb{R}$. Seeking a contradiction, assume that there is a limiting value ℓ for g at P. If this is so, then we take $\epsilon = 1/2$ and we can find a $\delta > 0$ such that $0 < |x - P| < \delta$ implies

$$|g(x) - \ell| < \epsilon = \frac{1}{2}. \tag{5.3.1}$$

If we take x to be rational then (5.3.1) says that

$$|1 - \ell| < \frac{1}{2}, \tag{5.3.2}$$

while if we take x irrational then (5.3.1) says that

$$|0 - \ell| < \frac{1}{2}. \tag{5.3.3}$$

But then the triangle inequality gives that

$$
\begin{aligned}
|1 - 0| &= |(1 - \ell) + (\ell - 0)| \\
&\le |1 - \ell| + |\ell - 0|,
\end{aligned}
$$

which by (5.3.2) and (5.3.3) is

$$
< 1 .
$$

This contradiction, that $1 < 1$, allows us to conclude that the limit does not exist at P. $\square$

Proposition 5.4 Let f be a function with domain E, and let P be an accumulation point of E. If $\lim_{x \to P} f(x) = \ell$ and $\lim_{x \to P} f(x) = m$, then $\ell = m$.

Proof: Let $\epsilon > 0$. Choose $\delta_1 > 0$ such that, if $x \in E$ and $0 < |x - P| < \delta_1$, then $|f(x) - \ell| < \epsilon/2$. Similarly choose $\delta_2 > 0$ such that, if $x \in E$ and $0 < |x - P| < \delta_2$, then $|f(x) - m| < \epsilon/2$. Define δ to be the minimum of δ_1 and δ_2. If $x \in E$ and $0 < |x - P| < \delta$, then the triangle inequality tells us that

$$
\begin{aligned}
|\ell - m| &= |(\ell - f(x)) + (f(x) - m)| \\
&\le |(\ell - f(x)| + |f(x) - m)| \\
&< \frac{\epsilon}{2} + \frac{\epsilon}{2} \\
&= \epsilon
\end{aligned}
$$

Since $|\ell - m| < \epsilon$ for every positive ϵ we conclude that $\ell = m$. That is the desired result. $\square$

The point of the last proposition is that if a limit is calculated by two different methods, then the same answer will result. While of primarily philosophical interest now, this will be important information later when we establish the existence of certain limits.

This is a good time to observe that the limits

$$
\lim_{x \to P} f(x)
$$

and

$$
\lim_{h \to 0} f(P + h)
$$

are equal in the sense that if one limit exists then so does the other and they both have the same value.

In order to facilitate checking that certain limits exist, we now record some elementary properties of the limit. This requires that we first recall how functions are combined.

Suppose that f and g are each functions which have domain E. We define the *sum* or *difference* of f and g to be the function

$$
(f \pm g)(x) = f(x) \pm g(x) ,
$$

the *product* of f and g to be the function

$$(f \cdot g)(x) = f(x) \cdot g(x),$$

and the quotient of f and g to be

$$\left(\frac{f}{g}\right)(x) = \frac{f(x)}{g(x)}.$$

Notice that the quotient is only defined at points x for which $g(x) \neq 0$. Now we have:

Theorem 5.5 (Elementary Properties of Limits of Functions) *Let f and g be functions with domain E and fix a point P that is an accumulation point of E. Assume that*

(i) $\lim_{x \to P} f(x) = \ell$

(ii) $\lim_{x \to P} g(x) = m$.

Then

(a) $\lim_{x \to P} (f \pm g)(x) = \ell \pm m$

(b) $\lim_{x \to P} (f \cdot g)(x) = \ell \cdot m$

(c) $\lim_{x \to P} (f/g)(x) = \ell/m$ *provided $m \neq 0$*.

Proof: We prove part **(b)**. Parts **(a)** and **(c)** are treated in the exercises.

Let $\epsilon > 0$. We may also assume that $\epsilon < 1$. Choose $\delta_1 > 0$ such that, if $x \in E$ and $0 < |x - P| < \delta_1$, then

$$|f(x) - \ell| < \frac{\epsilon}{2(|m| + 1)}.$$

Choose $\delta_2 > 0$ such that if $x \in E$ and $0 < |x - P| < \delta_2$ then

$$|g(x) - m| < \frac{\epsilon}{2(|\ell| + 1)}.$$

(Notice that this last inequality implies that $|g(x)| < |m| + |\epsilon|$.) Let δ be the minimum of δ_1 and δ_2. If $x \in E$ and $0 < |x - P| < \delta$ then

$$
\begin{aligned}
|f(x) \cdot g(x) - \ell \cdot m| &= |(f(x) - \ell) \cdot g(x) + (g(x) - m) \cdot \ell| \\
&\leq |(f(x) - \ell) \cdot g(x)| + |(g(x) - m) \cdot \ell| \\
&< \left(\frac{\epsilon}{2(|m| + 1)}\right) \cdot |g(x)| + \left(\frac{\epsilon}{2(|\ell| + 1)}\right) \cdot |\ell| \\
&\leq \left(\frac{\epsilon}{2(|m| + 1)}\right) \cdot (|m| + |\epsilon|) + \frac{\epsilon}{2} \\
&< \frac{\epsilon}{2} + \frac{\epsilon}{2} \\
&= \epsilon.
\end{aligned}
$$
$\qquad\square$

EXAMPLE 5.6 It is a simple matter to check that, if $f(x) = x$, then

$$\lim_{x \to P} f(x) = P$$

for every real P. (Indeed, for $\epsilon > 0$ we may take $\delta = \epsilon$.) Also, if $g(x) \equiv \alpha$ is the constant function taking value α, then

$$\lim_{x \to P} g(x) = \alpha \,.$$

It then follows from parts **(a)** and **(b)** of the theorem that, if $f(x)$ is any polynomial function, then

$$\lim_{x \to P} f(x) = f(P) \,.$$

Moreover, if $r(x)$ is any *rational function* (quotient of polynomials) then we may also use part **(c)** of the theorem to conclude that

$$\lim_{x \to P} r(x) = r(P)$$

for all points P at which the rational function $r(x)$ is defined.   □

EXAMPLE 5.7 If x is a small, positive real number then $0 < \sin x < x$. This is true because $\sin x$ is the nearest distance from the point $(\cos x, \sin x)$ to the x-axis while x is the distance from that point to the x-axis along an arc. If $\epsilon > 0$, then we set $\delta = \epsilon$. We conclude that if $0 < |x - 0| < \delta$ then

$$|\sin x - 0| < |x| < \delta = \epsilon \,.$$

Since $\sin(-x) = -\sin x$, the same result holds when x is a negative number with small absolute value. Therefore

$$\lim_{x \to 0} \sin x = 0 \,.$$

Since

$$\cos x = \sqrt{1 - \sin^2 x} \quad \text{for all} \ \ x \in [-\pi/2, \pi/2] \,,$$

we may conclude from the preceding theorem that

$$\lim_{x \to 0} \cos x = 1 \,.$$

Now fix any real number P. We have

$$
\begin{aligned}
\lim_{x \to P} \sin x &= \lim_{h \to 0} \sin(P + h) \\
&= \lim_{h \to 0} \left(\sin P \cos h + \cos P \sin h \right) \\
&= \sin P \cdot 1 + \cos P \cdot 0 \\
&= \sin P \,.
\end{aligned}
$$

We of course have used parts **(a)** and **(b)** of the theorem to commute the limit process with addition and multiplication. A similar argument shows that

$$\lim_{x \to P} \cos x = \cos P \,. \qquad \square$$

Remark 5.8 In the last example, we have used the definition of the sine function and the cosine function that you learned in calculus. In Chapter 8, when we learn about series of functions, we will learn a more rigorous method for treating the trigonometric functions.

We conclude by giving a characterization of the limit of a function using sequences.

Proposition 5.9 *Let f be a function with domain E and P be an accumulation point of E. Then*

$$\lim_{x \to P} f(x) = \ell \qquad (5.9.1)$$

if and only if, for any sequence $\{a_j\} \subseteq E \setminus \{P\}$ satisfying $\lim_{j \to \infty} a_j = P$, it holds that

$$\lim_{j \to \infty} f(a_j) = \ell \,. \qquad (5.9.2)$$

Proof: Assume that condition (5.9.1) fails. Then there is an $\epsilon > 0$ such that for no $\delta > 0$ is it the case that when $0 < |x - P| < \delta$ then $|f(x) - \ell| < \epsilon$. Thus, for each $\delta = 1/j$, we may choose a number $a_j \in E \setminus \{P\}$ with $0 < |a_j - P| < 1/j$ and $|f(a_j) - \ell| \geq \epsilon$. But then condition (5.9.2) fails for this sequence $\{a_j\}$.

If condition (5.9.2) fails then there is some sequence $\{a_j\}$ such that $\lim_{j \to \infty} a_j$
$= P$ but $\lim_{j \to \infty} f(a_j) \neq \ell$. This means that there is an $\epsilon > 0$ such that for infinitely many a_j it holds that $|f(a_j) - \ell| \geq \epsilon$. But then, no matter how small $\delta > 0$, there will be an a_j satisfying $0 < |a_j - P| < \delta$ (since $a_j \to P$) and $|f(a_j) - \ell| \geq \epsilon$. Thus (5.9.1) fails. $\qquad \square$

Exercises

1. Let f and g be functions on a set $A = (a, c) \cup (c, b)$ and assume that $f(x) \leq g(x)$ for all $x \in A$. Assuming that both limits exist, show that

$$\lim_{x \to c} f(x) \leq \lim_{x \to c} g(x) \,.$$

   Does the conclusion improve if we assume that $f(x) < g(x)$ for all $x \in A$?

2. Explain why it makes no sense to consider the limit of a function at an isolated point of the domain of the function.

3. Give a definition of limit using the concept of open set.

4. If $\lim_{x\to c} f(x) = \ell > 0$ then prove that there is a $\delta > 0$ so small that $|x - c| < \delta$ guarantees that $f(x) > \ell/2$.

5. Give an example of a function with domain $\mathbb{R}$ such that $\lim_{x\to c} f(x)$ exists at every point c but f is discontinuous at a dense set of points.

6. Prove that $\lim_{x\to P} f(x) = \lim_{h\to 0} f(P + h)$ whenever both expressions make sense.

7. Prove parts **(a)** and **(c)** of Theorem 5.5.

8. Give an example of a function $f : \mathbb{R} \to \mathbb{R}$ so that $\lim_{x\to c}$ does not exist for any $c \in \mathbb{R}$.

9. Discuss the limiting properties at the origin of the functions

$$f(x) = \begin{cases} x\sin(1/x) & \text{if} \quad x \neq 0 \\ 0 & \text{if} \quad x = 0 \,. \end{cases}$$

and

$$g(x) = \begin{cases} \sin(1/x) & \text{if} \quad x \neq 0 \\ 0 & \text{if} \quad x = 0 \,. \end{cases}$$

10. Show that, if f is an increasing or decreasing function, then f has a limit at "most" points. What does the word "most" mean in this context?

* **11.** Give an example of a function $f : \mathbb{R} \to \mathbb{R}$ so that $\lim_{x\to c} f(x)$ exists when c is irrational but does not exist when c is rational.

* **12.** Give a definition of limit using the concept of distance.

5.2 Continuous Functions

Definition 5.10 Let $E \subseteq \mathbb{R}$ be a set and let f be a real-valued function with domain E. Fix a point P which is in E and is also an accumulation point of E. We say that f is *continuous* at P if

$$\lim_{x\to P} f(x) = f(P)\,.$$

We learned from the penultimate example of Section 1 that polynomial functions are continuous at every real x. So are the transcendental functions $\sin x$ and $\cos x$ (see Example 5.7). A rational function is continuous at every point of its domain.

EXAMPLE 5.11 The function

$$h(x) = \begin{cases} \sin(1/x) & \text{if} \quad x \neq 0 \\ 1 & \text{if} \quad x = 0 \end{cases}$$
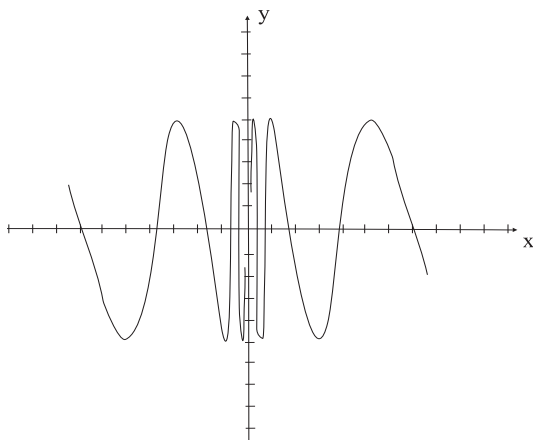
Figure 5.2: A function discontinuous at 0.

is discontinuous at 0. See Figure 5.2. The reason is that

$$\lim_{x \to 0} h(x)$$

does not exist. (Details of this assertion are left for you: notice that $h(1/(j\pi)) = 0$ while $h(2/[(4j+1)\pi]) = 1$ for $j = 1, 2, \ldots$.)

The function

$$k(x) = \begin{cases} x \cdot \sin(1/x) & \text{if} \quad x \neq 0 \\ 1 & \text{if} \quad x = 0 \end{cases}$$

is also discontinuous at $x = 0$. This time the limit $\lim_{x \to 0} k(x)$ exists (see Example 5.2), but the limit does not agree with $k(0)$.

However, the function

$$m(x) = \begin{cases} x \cdot \sin(1/x) & \text{if} \quad x \neq 0 \\ 0 & \text{if} \quad x = 0 \end{cases}$$

is continuous at $x = 0$ because the limit at 0 exists and agrees with the value of the function there. See Figure 5.3.   □

The arithmetic operations $+, -, \times$, and $\div$ preserve continuity (so long as we avoid division by zero). We now formulate this assertion as a theorem.

Theorem 5.12 *Let f and g be functions with domain E and let P be a point of E which is also an accumulation point of E. If f and g are continuous at P then so are $f \pm g, f \cdot g$, and (provided $g(P) \neq 0$) f/g.*

Proof: Apply Theorem 5.5 of Section 1.   □

Continuous functions may also be characterized using sequences:
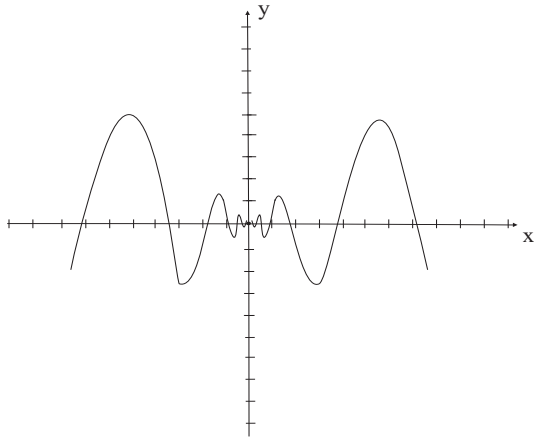
Figure 5.3: A function continuous at 0.
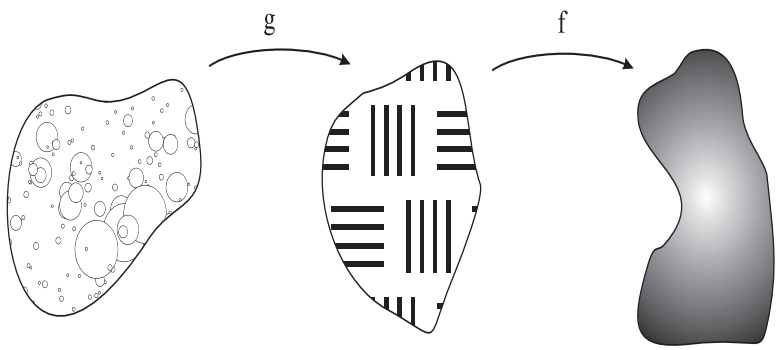
Figure 5.4: Composition of functions.

Proposition 5.13 *Let f be a function with domain E and fix $P \in E$ which is an accumulation point of E. The function f is continuous at P if and only if, for every sequence $\{a_j\} \subseteq E$ satisfying $\lim_{j \to \infty} a_j = P$, it holds that*

$$\lim_{j \to \infty} f(a_j) = f(P).$$

Proof: Apply Proposition 5.9 of Section 1. □

Recall that, if g is a function with domain D and range E, and if f is a function with domain E and range F, then the *composition* of f and g is

$$f \circ g(x) = f(g(x)).$$

See Figure 5.4.

Proposition 5.14 *Let g have domain D and range E and let f have domain E and range F. Let $P \in D$. Suppose that P is an accumulation point of D and $g(P)$ is an accumulation point of E. Assume that g is continuous at P and that f is continuous at $g(P)$. Then $f \circ g$ is continuous at P.*

Proof: Let $\{a_j\}$ be any sequence in D such that $\lim_{j\to\infty} a_j = P$. Then

$$\lim_{j\to\infty} f \circ g(a_j) = \lim_{j\to\infty} f(g(a_j)) = f\left(\lim_{j\to\infty} g(a_j)\right)$$

$$= f\left(g\left(\lim_{j\to\infty} a_j\right)\right) = f(g(P)) = f \circ g(P).$$

Now apply Proposition 5.9.   □

EXAMPLE 5.15 It is not the case that if

$$\lim_{x\to P} g(x) = \ell$$

and

$$\lim_{t\to\ell} f(t) = m$$

then

$$\lim_{x\to P} f \circ g(x) = m.$$

A counterexample is given by the functions

$$g(x) = 0$$

$$f(x) = \begin{cases} 2 & \text{if} \quad x \neq 0 \\ 5 & \text{if} \quad x = 0. \end{cases}$$

Notice that $\lim_{x\to 0} g(x) = 0, \lim_{t\to 0} f(t) = 2$, yet $\lim_{x\to 0} f \circ g(x) = 5$.

The additional hypothesis that f be continuous at ℓ is necessary in order to guarantee that the limit of the composition will behave as expected.   □

Next we explore the topological approach to the concept of continuity. Whereas the analytic approach that we have been discussing so far considers continuity one point at a time, the topological approach considers all points simultaneously. Let us call a function continuous if it is continuous at every point of its domain.

Definition 5.16 Let f be a function with domain E and let W be any set of real numbers. We define

$$f^{-1}(W) = \{x \in E : f(x) \in W\}.$$

We sometimes refer to $f^{-1}(W)$ as the *inverse image* of W under f.

Theorem 5.17 *Let f be a function with domain E. The function f is contin-uous if and only if the inverse image of any open set under f is the intersection of E with an open set.*

In particular, if E is open then f is continuous if and only if the inverse image of every open set under f is open.

Proof: Assume that f is continuous. Let $\mathcal{O}$ be any open set in $\mathbb{R}$ and let $P \in f^{-1}(\mathcal{O})$. Then, by definition, $f(P) \in \mathcal{O}$. Since $\mathcal{O}$ is open, there is an $\epsilon > 0$ such that the interval $(f(P) - \epsilon, f(P) + \epsilon)$ lies in $\mathcal{O}$. By the continuity of f we may select a $\delta > 0$ such that if $x \in E$ and $|x - P| < \delta$ then $|f(x) - f(P)| < \epsilon$. In other words, if $x \in E$ and $|x - P| < \delta$ then $f(x) \in \mathcal{O}$ or $x \in f^{-1}(\mathcal{O})$. Thus we have found an open interval $I = (P - \delta, P + \delta)$ about P whose intersection with E is contained in $f^{-1}(\mathcal{O})$. So $f^{-1}(\mathcal{O})$ is the intersection of E with an open set.

Conversely, suppose that for any open set $\mathcal{O} \subseteq \mathbb{R}$ we have that $f^{-1}(\mathcal{O})$ is the intersection of E with an open set. Fix $P \in E$. Choose $\epsilon > 0$. Then the interval $(f(P) - \epsilon, f(P) + \epsilon)$ is an open set. By hypothesis the set $f^{-1}((f(P) - \epsilon, f(P) + \epsilon))$ is the intersection of E with an open set. This set contains the point P. Thus there is a $\delta > 0$ such that

$$E \cap (P - \delta, P + \delta) \subseteq f^{-1}((f(P) - \epsilon, f(P) + \epsilon)).$$

But that just says that

$$f(E \cap (P - \delta, P + \delta)) \subseteq (f(P) - \epsilon, f(P) + \epsilon).$$

In other words, if $|x - P| < \delta$ and $x \in E$ then $|f(x) - f(P)| < \epsilon$. But that means that f is continuous at P. □

EXAMPLE 5.18 Since any open subset of the real numbers is a countable or finite disjoint union of intervals then—in order to check that the inverse image under a function f of every open set is open—it is enough to check that the inverse image of any open interval is open. This is frequently easy to do.

For example, if $f(x) = x^2$ then the inverse image of an open interval (a, b) is $(-\sqrt{b}, -\sqrt{a}) \cup (\sqrt{a}, \sqrt{b})$ if $a > 0$, is $(-\sqrt{b}, \sqrt{b})$ if $a \leq 0, b \geq 0$, and is $\emptyset$ if $a < b < 0$. Thus the function f is continuous.

Note that, by contrast, it is somewhat tedious to give an $\epsilon - \delta$ proof of the continuity of $f(x) = x^2$. □

Corollary 5.19 *Let f be a function with domain E. The function f is con-tinuous if and only if the inverse image of any closed set F under f is the intersection of E with some closed set.*

In particular, if E is closed then f is continuous if and only if the inverse image of any closed set F under f is closed.

Proof: It is enough to prove that

$$f^{-1}({}^c F) = {}^c \left(f^{-1}(F) \right).$$

We leave this assertion as an exercise for you. □

Exercises

1. Define the function

$$g(x) = \begin{cases} 0 & \text{if} \quad x \text{ is irrational} \\ x & \text{if} \quad x \text{ is rational} \end{cases}$$

   At which points x is g continuous? At which points is it discontinuous?

2. Let f be a continuous function whose domain contains an open interval (a, b). What form can $f((a, b))$ have? (**Hint:** There are just four possibilities.)

3. Explain why it would be foolish to define the concept of continuity at an isolated point.

 * 4. Let f be a continuous function on the open interval (a, b). Under what circumstances can f be extended to a continuous function on $[a, b]$?

5. Define an onto, continuous function from $\mathbb{R}^2$ to $\mathbb{R}$.

6. Define continuity using the notion of closed set.

7. Define the function $f(x)$ to equal 0 if x is irrational and to equal b if $x = a/b$ is a rational number in lowest terms. At which points is f continuous? At which points discontinuous?

8. The image of a compact set under a continuous function is compact (see the next section). But the image of a closed set need not be closed. Explain. The *inverse image* of a compact set under a continuous function need not be compact. Explain.

9. Give a careful proof of Corollary 5.19.

 * 10. See the next section for terminology. In particular, a function f on a set E is uniformly continuous if, given $\epsilon > 0$, there is a $\delta > 0$ such that $|f(s) - f(t)| < \epsilon$ whenever $|s - t| < \delta$.

   Let $0 < \alpha \leq 1$. A function f with domain E is said to satisfy a *Lipschitz condition* of order α if there is a constant $C > 0$ such that, for any $s, t \in E$, it holds that $|f(s) - f(t)| \leq C \cdot |s - t|^\alpha$. Prove that such a function must be uniformly continuous.

 * 11. See Exercise 10 for terminology. Is the composition of uniformly continuous functions uniformly continuous?
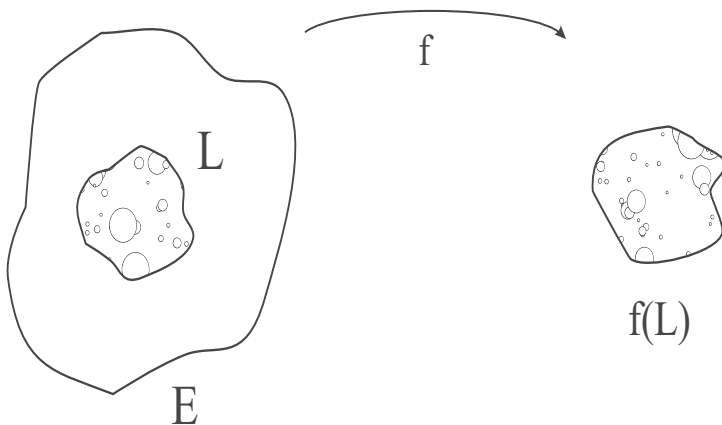
Figure 5.5: The image of the set L under the function f.

5.3 Topological Properties and Continuity

Recall that in Chapter 4 we learned a characterization of compact sets in terms of open covers. In Section 2 of the present chapter we learned a characterization of continuous functions in terms of inverse images of open sets. Thus it is not surprising that compact sets and continuous functions interact in a natural way. We explore this interaction in the present section.

Definition 5.20 Let f be a function with domain E and let L be a subset of E. We define
$$f(L) = \{f(x) : x \in L\}.$$
The set $f(L)$ is called the *image* of L under f. See Figure 5.5.

Theorem 5.21 *The image of a compact set under a continuous function is also compact.*

Proof: Let f be a continuous function with domain E and let K be a subset of E that is compact. Our job is to show that $f(K)$ is compact.

Let $\mathcal{C} = \{\mathcal{O}_\alpha\}$ be an open covering of $f(K)$. Since f is continuous we know that, for each α, the set $f^{-1}(\mathcal{O}_\alpha)$ is the intersection of E with an open set $\mathcal{U}_\alpha$. Let $\widehat{\mathcal{C}} = \{\mathcal{U}_\alpha\}_{\alpha \in A}$. Since $\mathcal{C}$ covers $f(K)$ it follows that $\widehat{\mathcal{C}}$ covers K. But K is compact; therefore (Theorem 4.37) there is a finite subcovering
$$\{\mathcal{U}_{\alpha_1}, \mathcal{U}_{\alpha_2}, \ldots \mathcal{U}_{\alpha_m}\}$$
of K. But then it follows that $f(\mathcal{U}_{\alpha_1} \cap E), \ldots, f(\mathcal{U}_{\alpha_m} \cap E)$ covers $f(K)$, hence
$$\mathcal{O}_{\alpha_1}, \mathcal{O}_{\alpha_2}, \ldots, \mathcal{O}_{\alpha_m}$$
covers $f(K)$.

We have taken an arbitrary open cover $\mathcal{C}$ for $f(K)$ and extracted from it a finite subcovering. It follows that $f(K)$ is compact.   □

It is not the case that the continuous image of a closed set is closed. For instance, take $f(x) = 1/(1 + x^2)$ and $E = \mathbb{R}$: the set E is closed and f is continuous but $f(E) = (0, 1]$ is not closed.

It is also not the case that the continuous image of a bounded set is bounded. As an example, take $f(x) = 1/x$ and $E = (0, 1)$. Then E is bounded and f continuous but $f(E) = (1, \infty)$ is unbounded.

However, the combined properties of closedness *and* boundedness (that is, compactness) are preserved. That is the content of the preceding theorem.

Corollary 5.22 *Let f be a continuous, real-valued function with compact domain $K \subseteq \mathbb{R}$. Then there is a number L such that*

$$|f(x)| \leq L$$

for all $x \in K$.

Proof: We know from the theorem that $f(K)$ is compact. By Theorem 4.29, we conclude that $f(K)$ is bounded. Thus there is a number L such that $|t| \leq L$ for all $t \in f(K)$. But that is just the assertion that we wish to prove.   □

In fact we can prove an important strengthening of the corollary. Since $f(K)$ is compact, it contains its supremum M and its infimum m. Therefore there must be a number $C \in K$ such that $f(C) = M$ and a number $c \in K$ such that $f(c) = m$. In other words, $f(c) \leq f(x) \leq f(C)$ for all $x \in K$. We summarize:

Theorem 5.23 *Let f be a continuous function on a compact set $K \subseteq \mathbb{R}$. Then there exist numbers c and C in K such that $f(c) \leq f(x) \leq f(C)$ for all $x \in K$. We call c an absolute minimum for f on K and C an absolute maximum for f on K. We call $f(c)$ the absolute minimum value for f on K and $f(C)$ the absolute maximum value for f on K.*

Notice that, in the last theorem, the location of the absolute maximum and absolute minimum need not be unique. For instance, the function $\sin x$ on the compact interval $[0, 4\pi]$ has an absolute minimum at $3\pi/2$ and $7\pi/2$. It has an absolute maximum at $\pi/2$ and at $5\pi/2$.

Now we define a refined type of continuity called "uniform continuity." We shall learn that this new notion of continuous function arises naturally for a continuous function on a compact set. It will also play an important role in our later studies, especially in the context of the integral.

Definition 5.24 *Let f be a function with domain $E \subseteq \mathbb{R}$. We say that f is uniformly continuous on E if, for each $\epsilon > 0$, there is a $\delta > 0$ such that, whenever $s, t \in E$ and $|s - t| < \delta$, then $|f(s) - f(t)| < \epsilon$.*

Observe that "uniform continuity" differs from "continuity" in that it treats all points of the domain simultaneously: the $\delta > 0$ that is chosen is independent of the points $s, t \in E$. This difference is highlighted by the next two examples.

EXAMPLE 5.25 Suppose that a function $f : \mathbb{R} \to \mathbb{R}$ satisfies the condition

$$|f(s) - f(t)| \leq C \cdot |s - t|, \tag{5.25.1}$$

where C is some positive constant. This is called a *Lipschitz condition*, and it arises frequently in analysis. Let $\epsilon > 0$ and set $\delta = \epsilon/C$. If $|x - y| < \delta$ then, by (5.25.1),

$$|f(x) - f(y)| \leq C \cdot |x - y| < C \cdot \delta = C \cdot \frac{\epsilon}{C} = \epsilon.$$

It follows that f is uniformly continuous. □

EXAMPLE 5.26 Consider the function $f(x) = x^2$. Fix a point $P \in \mathbb{R}, P > 0$, and let $\epsilon > 0$. In order to guarantee that $|f(x) - f(P)| < \epsilon$ we must have (for $x > 0$)

$$|x^2 - P^2| < \epsilon$$

or

$$|x - P| < \frac{\epsilon}{x + P}.$$

Since x will range over a neighborhood of P, we see that the required δ in the definition of continuity cannot be larger than $\epsilon/(2P)$. In fact the choice $|x - P| < \delta = \epsilon/(2P + 1)$ will do the job.

Put in slightly different words, let $\epsilon = 1$. Then $|f(j + 1/j) - f(j)| > \epsilon = 1$ for any j. Thus, for this ϵ, we may not take δ to be $1/j$ for any j. So no uniform δ exists.

Thus the choice of δ depends not only on ϵ (which we have come to expect) but also on P. In particular, f is not uniformly continuous on $\mathbb{R}$. This is a quantitative reflection of the fact that the graph of f becomes ever steeper as the variable x moves to the right.

Notice that the same calculation shows that the function f with restricted domain $[a, b], 0 < a < b < \infty$, *is* uniformly continuous. That is because, when the function is restricted to $[a, b]$, its slope does not become arbitrarily large. See Figure 5.6. □

Now the main result about uniform continuity is the following:

Theorem 5.27 *Let f be a continuous function with compact domain K. Then f is uniformly continuous on K.*

Proof: Pick $\epsilon > 0$. By the definition of continuity there is for each point $x \in K$ a number $\delta_x > 0$ such that if $|x - t| < \delta_x$ then $|f(t) - f(x)| < \epsilon/2$. The intervals $I_x = (x - \delta_x/2, x + \delta_x/2)$ form an open covering of K. Since K is compact, we may therefore (by Theorem 4.34) extract a finite subcovering

$$I_{x_1}, \dots I_{x_m}.$$
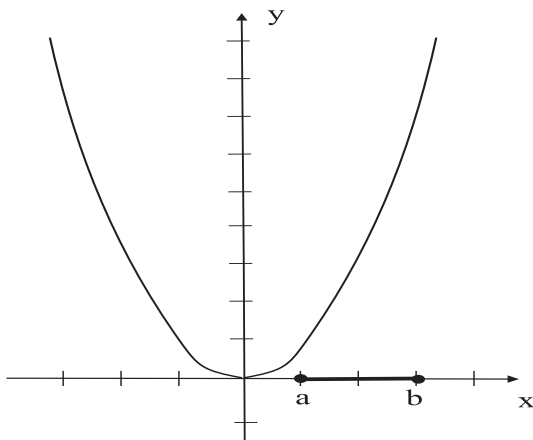
Figure 5.6: Uniform continuity on the interval $[a, b]$.

Now let $\delta = \min\{\delta_{x_1}/2, \ldots, \delta_{x_m}/2\} > 0$. If $s, t \in K$ and $|s - t| < \delta$ then $s \in I_{x_j}$ for some $1 \leq j \leq m$. It follows that

$$|s - x_j| < \delta_{x_j}/2$$

and

$$|t - x_j| \leq |t - s| + |s - x_j| < \delta + \delta_{x_j}/2 \leq \delta_{x_j}/2 + \delta_{x_j}/2 = \delta_{x_j}.$$

We know that

$$|f(s) - f(t)| \leq |f(s) - f(x_j)| + |f(x_j) - f(t)|.$$

But since each of s and t is within δ_{x_j} of x_j we may conclude that the last line is less than

$$\frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.$$

Notice that our choice of δ does not depend on s and t (indeed, we chose δ *before* we chose s and t). We conclude that f is uniformly continuous. □

Remark 5.28 Where in the proof did the compactness play a role? We defined δ to be the minimum of $\delta_{x_1}, \ldots \delta_{x_m}$. In order to guarantee that δ be *positive* it is crucial that we be taking the minimum of *finitely many* positive numbers. So we needed a *finite* subcovering.

EXAMPLE 5.29 The function $f(x) = \sin(1/x)$ is continuous on the domain $E = (0, \infty)$ since it is the composition of continuous functions (refer again to Figure 5.2). However, it is not uniformly continuous since

$$\left| f\left(\frac{1}{2j\pi}\right) - f\left(\frac{1}{\frac{(4j+1)\pi}{2}}\right) \right| = 1$$

for $j = 1, 2, \ldots$. Thus, even though the arguments are becoming arbitrarily close together, the images of these arguments remain bounded apart. We conclude that f cannot be uniformly continuous. See Figure 5.2.

However, if f is considered as a function on any interval of the form $[a, b], 0 < a < b < \infty$, then the preceding theorem tells us that the function f is uniformly continuous. ☐

As an exercise, you should check that

$$g(x) = \begin{cases} x \sin(1/x) & \text{if} \quad x \neq 0 \\ 0 & \text{if} \quad x = 0 \end{cases}$$

is uniformly continuous on any interval of the form $[-N, N]$. See Figure 5.3.

Next we show that continuous functions preserve connectedness.

Theorem 5.30 *Let f be a continuous function with domain an open interval I. Suppose that L is a connected subset of I. Then $f(L)$ is connected.*

Proof: Suppose to the contrary that there are open sets U and V such that

$$U \cap f(L) \neq \emptyset, V \cap f(L) \neq \emptyset,$$

$$(U \cap f(L)) \cap (V \cap f(L)) = \emptyset,$$

and

$$f(L) = (U \cap f(L)) \cup (V \cap f(L)) .$$

Since f is continuous, $f^{-1}(U)$ and $f^{-1}(V)$ are open. They each have nonempty intersection with L since $U \cap f(L)$ and $V \cap f(L)$ are nonempty. By the definition of f^{-1}, they are disjoint. And since $U \cup V$ contains $f(L)$ it follows, by definition, that $f^{-1}(U) \cup f^{-1}(V)$ contains L. But this shows that L is disconnected, and that is a contradiction. ☐

Corollary 5.31 (The Intermediate Value Theorem) *Let f be a continuous function whose domain contains the interval $[a, b]$. Let γ be a number that lies between $f(a)$ and $f(b)$. Then there is a number c between a and b such that $f(c) = \gamma$. Refer to Figure 5.7.*

Proof: The set $[a, b]$ is connected. Therefore $f([a, b])$ is connected. But $f([a, b])$ contains the points $f(a)$ and $f(b)$. By connectivity, $f([a, b])$ must contain the interval that has $f(a)$ and $f(b)$ as endpoints. In particular, $f([a, b])$ must contain any number γ that lies between $f(a)$ and $f(b)$. But this just says that there is a number c lying between a and b such that $f(c) = \gamma$. That is the desired conclusion. ☐
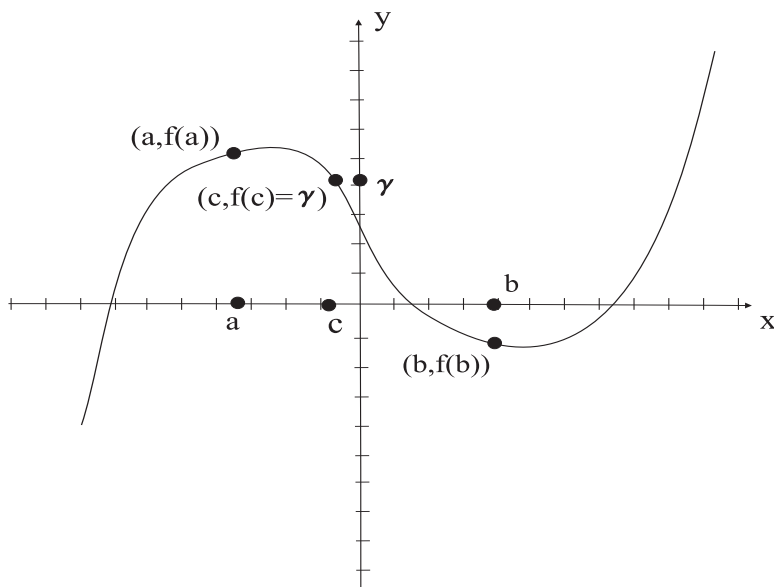
Figure 5.7: The Intermediate Value Theorem.

EXAMPLE 5.32 Let f be a continuous function with domain the interval $[0, 1]$ and range the interval $[0, 1]$. We claim that f has a fixed point, that is, a point p such that $f(p) = p$.

To see this, suppose not. Consider the function $g(x) = f(x) - x$. Since f has no fixed point, $f(0) > 0$ so $g(0) > 0$. Also $f(1) < 1$ so that $g(1) < 0$. By the Intermediate Value Theorem, it follows that there is a point p at which $g(p) = 0$. But that means that $f(p) = p$.   □

Exercises

1. If f is continuous on $[0, 1]$ and if $f(x)$ is positive for each rational x, then does it follow that f is positive at all x?

2. Give an example of a continuous function f and a connected set E such that $f^{-1}(E)$ is not connected. Is there a condition you can add that will force $f^{-1}(E)$ to be connected?

3. Give an example of a continuous function f and an open set U so that $f(U)$ is not open.

4. Let S be any subset of $\mathbb{R}$. Define the function

$$f(x) = \inf\{|x - s| : s \in S\}.$$

[We think of $f(x)$ as the distance of x to S.] Prove that f is uniformly continuous.

5. Let f be any function whose domain and range is the entire real line. If A and B are disjoint sets does it follow that $f(A)$ and $f(B)$ are disjoint sets? If C and D are disjoint sets does it follow that $f^{-1}(C)$ and $f^{-1}(D)$ are disjoint?

6. Let f be any function whose domain is the entire real line. If A and B are sets then is $f(A \cup B) = f(A) \cup f(B)$? If C and D are sets then is $f^{-1}(C \cup D) = f^{-1}(C) \cup f^{-1}(D)$? What is the answer to these questions if we replace $\cup$ by $\cap$?

7. We know that the continuous image of a connected set (i.e., an interval) is also a connected set (another interval). Suppose now that A is the union of k disjoint intervals and that f is a continuous function. What can you say about the set $f(A)$?

8. A function f with domain A and range B is called a *homeomorphism* if it is one-to-one, onto, continuous, and has a continuous inverse. If such an f exists then we say that A and B are *homeomorphic*. Which sets of reals are homeomorphic to the open unit interval $(0, 1)$? Which sets of reals are homeomorphic to the closed unit interval $[0, 1]$?

9. Let f be a continuous function with domain $[0, 1]$ and range $[0, 1]$. We know from Example 5.32 that there exists a point $P \in [0, 1]$ such that $f(P) = P$. Prove that this result is false if the domain and range of the function are both $(0, 1)$.

10. Let f be a continuous function and let $\{a_j\}$ be a Cauchy sequence in the domain of f. Does it follow that $\{f(a_j)\}$ is a Cauchy sequence? What if we assume instead that f is uniformly continuous?

11. Let E and F be disjoint closed sets of real numbers. Prove that there is a continuous function f with domain the real numbers such that $\{x : f(x) = 0\} = E$ and $\{x : f(x) = 1\} = F$.

12. If K and L are sets then define

$$K + L = \{k + \ell : k \in K \text{ and } \ell \in L\}.$$

If K and L are compact then prove that $K + L$ is compact. If K and L are merely closed, does it follow that $K + L$ is closed?

* 13. A function f from an interval (a, b) to an interval (c, d) is called *proper* if, for any compact set $K \subseteq (c, d)$, it holds that $f^{-1}(K)$ is compact. Prove that if f is proper then either

$$\lim_{x \to a^+} f(x) = c \quad \text{or} \quad \lim_{x \to a^+} f(x) = d.$$

Likewise prove that either

$$\lim_{x \to b^-} f(x) = c \quad \text{or} \quad \lim_{x \to b^-} f(x) = d.$$

*   **14.** Let E be any closed set of real numbers. Prove that there is a continuous function f with domain $\mathbb{R}$ such that $\{x : f(x) = 0\} = E$.

*   **15.** Prove that the function $f(x) = \cos x$ can be written, on the interval $(0, 2\pi)$, as the difference of two increasing functions.

5.4   Classifying Discontinuities and Monotonicity

We begin by refining our notion of limit:

Definition 5.33 Fix $P \in \mathbb{R}$. Let f be a function with domain E. Suppose that P is a limit point of $E \cap [P - 1, P)$. We say that f has *left limit* ℓ at P, and write

$$\lim_{x \to P-} f(x) = \ell$$

if, for every $\epsilon > 0$, there is a $\delta > 0$ such that, whenever $x \in E$ and $P - \delta < x < P$, then it holds that

$$|f(x) - \ell| < \epsilon.$$

Now suppose that P is a limit point of $E \cap (P, P + 1]$. We say that f has *right limit* m at P, and write

$$\lim_{x \to P+} f(x) = m$$

if, for every $\epsilon > 0$, there is a $\delta > 0$ such that, whenever $x \in E$ and $P < x < P + \delta$, then it holds that

$$|f(x) - m| < \epsilon.$$

This definition simply formalizes the notion of either letting x tend to P from the left only or from the right only.

EXAMPLE 5.34 Let

$$f(x) = \begin{cases} 0 & \text{if} \quad x \leq 0 \\ \sin(1/x) & \text{if} \quad x > 0. \end{cases}$$

Then $\lim_{x \to 0-} f(x) = 0$ and $\lim_{x \to 0+} f(x)$ does not exist.   □

Definition 5.35 Fix $P \in \mathbb{R}$. Let f be a function with domain E. Suppose that P is a limit point of $E \cap [P - 1, P)$ and that P is an element of E. We say that f is *left continuous* at P if

$$\lim_{x \to P-} f(x) = f(P).$$

Likewise, in case P is a limit point of $E \cap (P, P + 1]$ and is also an element of E, we say that f is *right continuous* at P if

$$\lim_{x \to P+} f(x) = f(P).$$

discontinuity of the first kind   discontinuity of the second kind
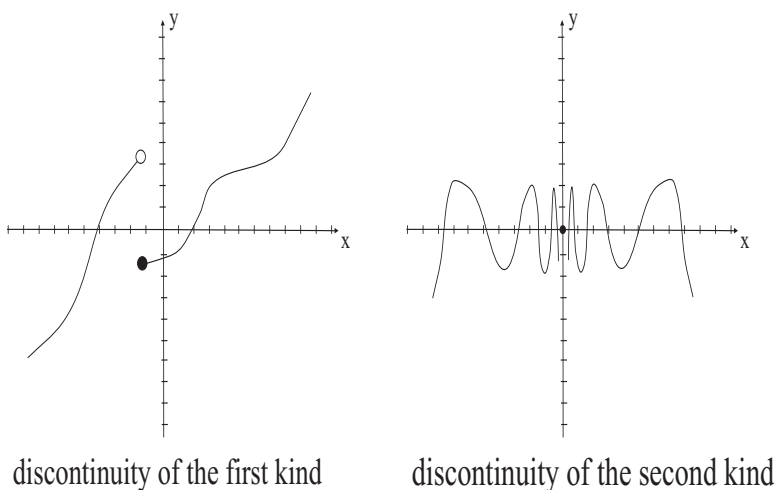
Figure 5.8: Discontinuities of the first and second kind.

EXAMPLE 5.36 Define

$$f(x) = \begin{cases} 1 & \text{if} & x < 0 \\ 0 & \text{if} & x = 0 \\ x\sin(1/x) & \text{if} & x > 0. \end{cases}$$

Then f is right continuous at 0 but f is not left continuous at 0. $\qquad\square$

Let f be a function with domain E. Let P in E and assume that f is discontinuous at P. There are two ways in which this discontinuity can occur:

I. If $\lim_{x\to P-} f(x)$ and $\lim_{x\to P+} f(x)$ both exist but either do not equal each other or do not equal $f(P)$ then we say that f has a *discontinuity of the first kind* (or sometimes a *simple discontinuity*) at P.

II. If either $\lim_{x\to P-}$ does not exist or $\lim_{x\to P+}$ does not exist then we say that f has a *discontinuity of the second kind* at P.

Refer to Figure 5.8.

EXAMPLE 5.37 Define

$$f(x) = \begin{cases} \sin(1/x) & \text{if} \quad x \neq 0 \\ 0 & \text{if} \quad x = 0 \end{cases}$$

$$g(x) = \begin{cases} 1 & \text{if} \quad x > 0 \\ 0 & \text{if} \quad x = 0 \\ -1 & \text{if} \quad x < 0 \end{cases}$$

monotone increasing function   monotone decreasing function
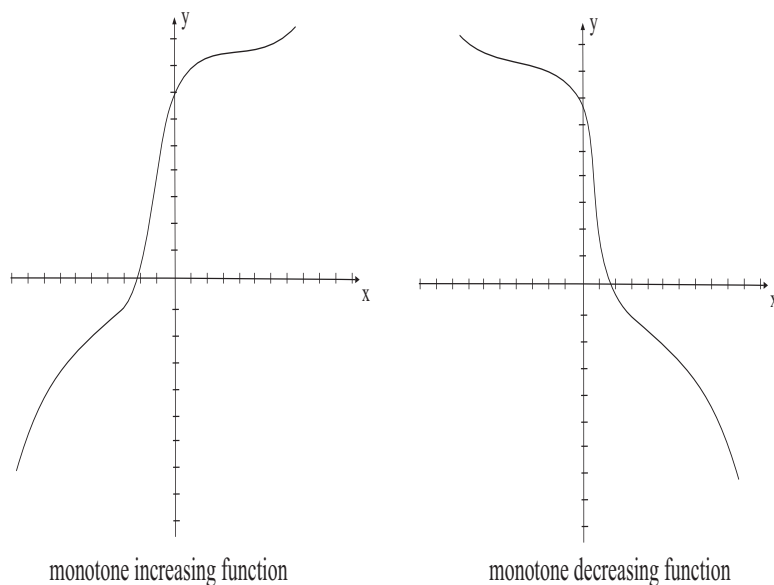
Figure 5.9: Increasing and decreasing functions.

$$h(x) = \left\{ \begin{array}{ll} 1 & \text{if } x \text{ is irrational} \\ 0 & \text{if } x \text{ is rational} \end{array} \right.$$

Then f has a discontinuity of the second kind at 0 while g has a discontinuity of the first kind at 0. The function h has a discontinuity of the second kind at every point. □

Definition 5.38 Let f be a function whose domain contains an open interval (a, b). We say that f is *increasing* on (a, b) if, whenever $a < s < t < b$, it holds that $f(s) \leq f(t)$. We say that f is *decreasing* on (a, b) if, whenever $a < s < t < b$, it holds that $f(s) \geq f(t)$. See Figure 5.9.

If a function is either increasing or decreasing then we call it *monotone* or *monotonic*. Compare with the definition of monotonic sequences in Section 2.1.

As with sequences, the word "monotonic" is superfluous in many contexts. But its use is traditional and occasionally convenient.

Proposition 5.39 *Let f be a monotonic function on an open interval (a, b). Then all of the discontinuities of f are of the first kind.*

Proof: It is enough to show that, for each $P \in (a, b)$, the limits

$$\lim_{x \to P^-} f(x)$$

and
$$\lim_{x \to P+} f(x)$$
exist.

Let us first assume that f is monotonically increasing. Fix $P \in (a, b)$. If $a < s < P$ then $f(s) \leq f(P)$. Therefore $S = \{f(s) : a < s < P\}$ is bounded above. Let M be the least upper bound of S. Pick $\epsilon > 0$. By definition of least upper bound there must be an $f(s) \in S$ such that $|f(s) - M| < \epsilon$. Let $\delta = |P - s|$. If $P - \delta < t < P$ then $s < t < P$ and $f(s) \leq f(t) \leq M$ or $|f(t) - M| < \epsilon$. Thus $\lim_{x \to P-} f(x)$ exists and equals M.

If we set m equal to the infimum of the set $T = \{f(t) : P < t < b\}$ then a similar argument shows that $\lim_{x \to P+} f(x)$ exists and equals m. That completes the proof.

The argument in case f is monotonically decreasing is just the same, and we omit the details. $\square$

Corollary 5.40 *Let f be a monotonic function on an interval (a, b). Then f has at most countably many discontinuities.*

Proof: Assume for simplicity that f is monotonically increasing. If P is a discontinuity then the proposition tells us that
$$\lim_{x \to P-} f(x) < \lim_{x \to P+} f(x).$$

Therefore there is a rational number q_P between $\lim_{x \to P-} f(x)$ and $\lim_{x \to P+} f(x)$. Notice that different discontinuities will have different rational numbers associated to them because if $\widehat{P}$ is another discontinuity and, say, $\widehat{P} < P$ then
$$\lim_{x \to \widehat{P}-} f(x) < q_{\widehat{P}} < \lim_{x \to \widehat{P}+} f(x) \leq \lim_{x \to P-} f(x) < q_P < \lim_{x \to P+} f(x).$$

Thus we have exhibited a one-to-one function from the set of discontinuities of f into the set of rational numbers. It follows that the set of discontinuities is countable.

The argument in case f is monotonically decreasing is just the same, and we omit the details. $\square$

A continuous function f has the property that the inverse image under f of any open set is open. However, it is not in general true that the *image* under f of any open set is open. A counterexample is the function $f(x) = x^2$ and the open set $\mathcal{O} = (-1, 1)$ whose image under f is $[0, 1)$.

EXAMPLE 5.41 Consider the greatest integer function $f(x) = [x]$. This means that $f(x)$ equals the greatest integer which is less than or equal to x. Then f is monotone increasing, and its discontinuities are of the first kind and are at the integers. This example illustrates Proposition 5.39 and Corollary 5.40. $\square$
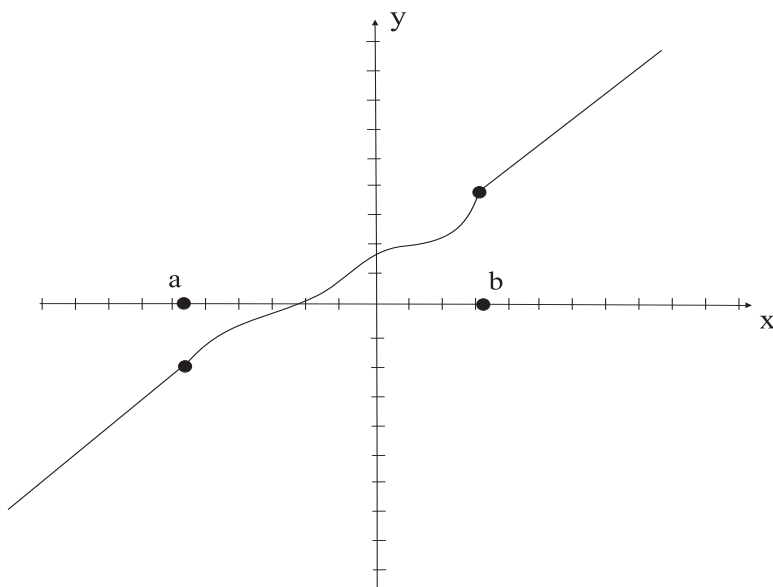
Figure 5.10: A strictly monotonically increasing function.

Definition 5.42 Suppose that f is a function on (a, b) such that $a < s < t < b$ implies $f(s) < f(t)$. Such a function is called *strictly increasing* (*strictly decreasing* functions are defined similarly). We refer to such functions as *strictly monotone*.

It is clear that a strictly increasing (resp. decreasing) function is one-to-one, hence has an inverse. Now we prove:

Theorem 5.43 *Let f be a strictly monotone, continuous function with domain $[a, b]$. Then f^{-1} exists and is continuous.*

Proof: Assume without loss of generality that f is strictly monotone *increasing*. Let us extend f to the entire real line by defining

$$ f(x) = \begin{cases} (x - a) + f(a) & \text{if} \quad x < a \\ \text{as given} & \text{if} \quad a \le x \le b \\ (x - b) + f(b) & \text{if} \quad x > b. \end{cases} $$

See Figure 5.10. Then it is easy to see that this extended version of f is still continuous and is strictly monotone increasing on all of $\mathbb{R}$.

That f^{-1} exists has already been discussed. The extended function f takes any open interval (c, d) to the open interval $(f(c), f(d))$. Since any open set is a union of open intervals, we see that f takes any open set to an open set. In other words, $[f^{-1}]^{-1}$ takes open sets to open sets. But this just says that f^{-1} is continuous.

Since the inverse of the extended function f is continuous, then so is the inverse of the original function f. That completes the proof. □

EXAMPLE 5.44 Consider the function

$$f(x) = e^x .$$

It is strictly inreasing on the entire real line, so it has an inverse. Its inverse is in fact the natural logarithm function $\ln x$.

Now take a look at the function

$$g(x) = \begin{cases} x & \text{if} & x \le -1 \\ -1 & \text{if} & -1 < x < 1 \\ x - 2 & \text{if} & 1 \le x . \end{cases}$$

This is a monotone increasing function, but it is *not* strictly increasing. And it has no inverse because it is not one-to-one. □

Exercises

1. Let A be any left-to-right ordered, countable subset of the reals. Assume that A has no accumulation points. In particular, $A = \{a_j\}$ and $a_j \to \infty$. Construct an increasing function whose set of points of discontinuity is precisely the set A. Explain why this is, in general, impossible for an uncountable set A.

2. Give an example of two functions, discontinuous at $x = 0$, whose sum *is* continuous at $x = 0$. Give an example of two such functions whose product is continuous at $x = 0$. How does the problem change if we replace "product" by "quotient"?

3. Let f be a function with domain $\mathbb{R}$. If $f^2(x) = f(x) \cdot f(x)$ is continuous, then does it follow that f is continuous? If $f^3(x) = f(x) \cdot f(x) \cdot f(x)$ is continuous, then does it follow that f is continuous?

4. Fix an interval (a, b). Is the collection of increasing functions on (a, b) closed under $+, -, \times$, or $\div$?

5. Let f be a continuous function whose domain contains a closed, bounded interval $[a, b]$. What topological properties does $f([a, b])$ possess? Is this set necessarily an interval?

6. Refer to Exercise **9** of Section 5.3 for terminology. Show that there is no homeomorphism from the real line to the interval $[0, 1)$.

7. Let f be a function with domain $\mathbb{R}$. Prove that the set of discontinuities of the first kind for f is countable. (**Hint:** If the left and right limits at a point disagree then you can slip a rational number between them.)

8. Let $a_1 < a_2 < \cdots$ with the a_j increasing to infinity and the a_j having no finite accumulation points. Give an example of a function with a discontinuity of the second kind at each a_j and no other discontinuities.

* **9.** Let $I \subseteq \mathbb{R}$ be an open interval and $f : I \to \mathbb{R}$ a function. We say that f is *convex* if whenever $\alpha, \beta \in I$ and $0 \leq t \leq 1$ then

$$f((1-t)\alpha + t\beta) \leq (1-t)f(\alpha) + tf(\beta) \,.$$

Prove that a convex function must be continuous. What does this definition of convex function have to do with the notion of "concave up" that you learned in calculus?

* **10.** Refer to Exercise **9** for terminology. What can you say about differentiability of a convex function?

* **11.** TRUE or FALSE: If f is a continuous function with domain and range the real numbers and which is both one-to-one and onto, then f must be either increasing or decreasing. Does your answer change if we assume that f is continuously differentiable (see the next chapter for this terminology)?

Chapter 6

Differentiation of Functions

6.1   The Concept of Derivative

Let f be a function with domain an open interval I. If $x \in I$ then the quantity

$$\frac{f(t) - f(x)}{t - x}$$

measures the slope of the chord of the graph of f that connects the points $(x, f(x))$ and $(t, f(t))$. See Figure 6.1. If we let $t \to x$ then the limit of the quantity represented by this "Newton quotient" should represent the slope of the graph *at the point* x. These considerations motivate the definition of the derivative:

Definition 6.1 If f is a function with domain an open interval I and if $x \in I$ then the limit

$$\lim_{t \to x} \frac{f(t) - f(x)}{t - x},$$

when it exists, is called the *derivative* of f at x. See Figure 6.2. If the derivative of f at x exists then we say that f is *differentiable* at x. If f is differentiable at every $x \in I$ then we say that f is *differentiable on* I.

We write the derivative of f at x either as

$$f'(x) \quad \text{or} \quad \frac{d}{dx}f \quad \text{or} \quad \frac{df}{dx} \quad \text{or} \quad \dot{f}.$$

We begin our discussion of the derivative by establishing some basic properties and relating the notion of derivative to continuity.

Lemma 6.2 If f is differentiable at a point x then f is continuous at x. In particular, $\lim_{t \to x} f(t) = f(x)$.
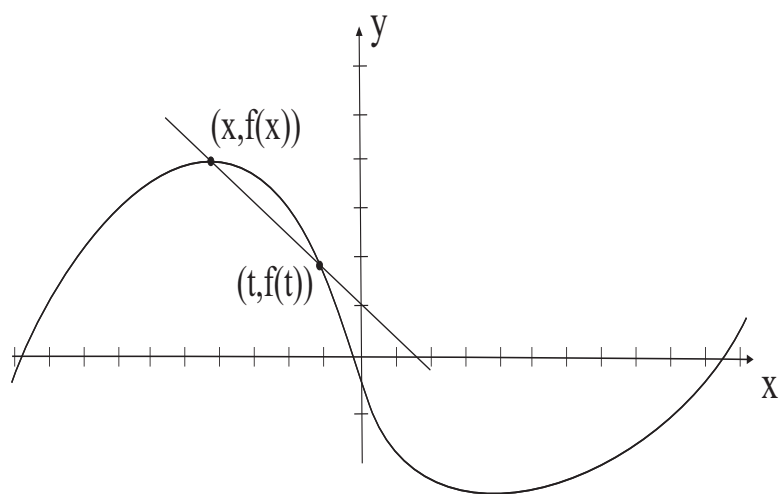
Figure 6.1: The Newton quotient.
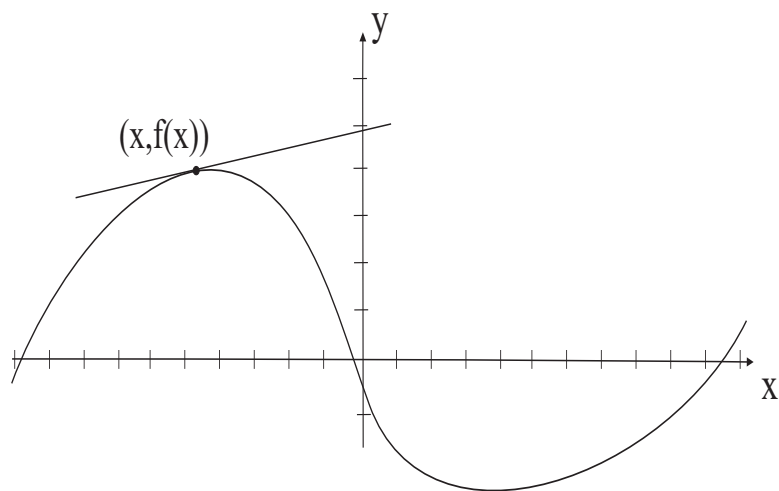
Figure 6.2: The derivative.

Proof: We use Theorem 5.5(**b**) about limits to see that

$$
\begin{aligned}
\lim_{t \to x} (f(t) - f(x)) &= \lim_{t \to x} \left((t - x) \cdot \frac{f(t) - f(x)}{t - x} \right) \\
&= \lim_{t \to x} (t - x) \cdot \lim_{t \to x} \frac{f(t) - f(x)}{t - x} \\
&= 0 \cdot f'(x) \\
&= 0.
\end{aligned}
$$

Therefore $\lim_{t \to x} f(t) = f(x)$ and f is continuous at x. $\square$

EXAMPLE 6.3 All differentiable functions are continuous: differentiability is a stronger property than continuity. Observe that the function $f(x) = |x|$ is continuous at every x but is not differentiable at 0. So continuity does not imply differentiability. Details appear in Example 6.5 below. $\square$

Theorem 6.4 *Assume that f and g are functions with domain an open interval I and that f and g are differentiable at $x \in I$. Then $f \pm g$, $f \cdot g$, and f/g are differentiable at x (for f/g we assume that $g(x) \neq 0$). Moreover*

(a) $(f \pm g)'(x) = f'(x) \pm g'(x);$

(b) $(f \cdot g)'(x) = f'(x) \cdot g(x) + f(x) \cdot g'(x);$

(c) $\left(\dfrac{f}{g} \right)'(x) = \dfrac{g(x) \cdot f'(x) - f(x) \cdot g'(x)}{g^2(x)}.$

Proof: Assertion (**a**) is easy and we leave it as an exercise for you.
   For (**b**), we write

$$
\begin{aligned}
\lim_{t \to x} \frac{(f \cdot g)(t) - (f \cdot g)(x)}{t - x} &= \lim_{t \to x} \left(\frac{(f(t) - f(x)) \cdot g(t)}{t - x} \right. \\
&\qquad \left. + \frac{(g(t) - g(x)) \cdot f(x)}{t - x} \right) \\
&= \lim_{t \to x} \left(\frac{(f(t) - f(x)) \cdot g(t)}{t - x} \right) \\
&\qquad + \lim_{t \to x} \left(\frac{(g(t) - g(x)) \cdot f(x)}{t - x} \right) \\
&= \lim_{t \to x} \left(\frac{(f(t) - f(x))}{t - x} \right) \cdot \left(\lim_{t \to x} g(t) \right) \\
&\qquad + \lim_{t \to x} \left(\frac{(g(t) - g(x))}{t - x} \right) \cdot \left(\lim_{t \to x} f(x) \right),
\end{aligned}
$$

where we have used Theorem 5.5 about limits. Now the first limit is the derivative of f at x, while the third limit is the derivative of g at x. Also notice that the limit of $g(t)$ equals $g(x)$ by the lemma. The result is that the last line equals

$$f'(x) \cdot g(x) + g'(x) \cdot f(x),$$

as desired.

To prove **(c)**, write

$$\lim_{t \to x} \frac{(f/g)(t) - (f/g)(x)}{t - x} = \lim_{t \to x} \frac{1}{g(t) \cdot g(x)} \left(\frac{f(t) - f(x)}{t - x} \cdot g(x) \right.$$
$$\left. - \frac{g(t) - g(x)}{t - x} \cdot f(x) \right).$$

The proof is now completed by using Theorem 5.5 about limits to evaluate the individual limits in this expression. $\square$

EXAMPLE 6.5 That $f(x) = x$ is differentiable follows from

$$\lim_{t \to x} \frac{t - x}{t - x} = 1.$$

Any constant function is differentiable (with derivative identically zero) by a similar argument. It follows from the theorem that any polynomial function is differentiable.

On the other hand, the continuous function $f(x) = |x|$ is *not* differentiable at the point $x = 0$. This is so because

$$\lim_{t \to 0^-} \frac{|t| - |0|}{t - x} = \lim_{t \to 0^-} \frac{-t - 0}{t - 0} = -1$$

while

$$\lim_{t \to 0^+} \frac{|t| - |0|}{t - x} = \lim_{t \to 0^+} \frac{t - 0}{t - 0} = 1.$$

So the required limit does not exist. $\square$

Since the subject of differential calculus is concerned with learning uses of the derivative, it concentrates on functions which *are* differentiable. One comes away from the subject with the impression that most functions are differentiable except at a few isolated points—as is the case with the function $f(x) = |x|$. Indeed this was what the mathematicians of the nineteenth century thought. Therefore it came as a shock when Karl Weierstrass produced a continuous function that is not differentiable at *any point*. In a sense that can be made precise, *most* continuous functions are of this nature: their graphs "wiggle" so much that they cannot have a tangent line at any point. Now we turn to an elegant variant of the example of Weierstrass that is due to B. L. van der Waerden (1903–1996).
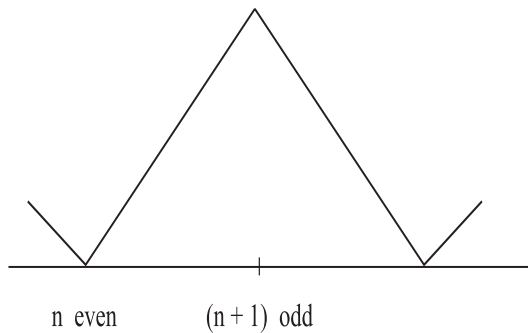
Figure 6.3: The van der Waerden example.

Theorem 6.6 *Define a function ψ with domain $\mathbb{R}$ by the rule*

$$\psi(x) = \begin{cases} x - n & \text{if} \quad n \leq x < n+1 \text{ and } n \text{ is even} \\ n+1-x & \text{if} \quad n \leq x < n+1 \text{ and } n \text{ is odd} \end{cases}$$

for every integer n. The graph of this function is exhibited in Figure 6.3. Then the function

$$f(x) = \sum_{j=1}^{\infty} \left(\frac{3}{4}\right)^j \psi\left(4^j x\right)$$

is continuous at every real x and differentiable at no real x.

Proof: Since we have not yet discussed series of functions, we take a moment to understand the definition of f. Fix a real x. Notice that $0 \leq \psi(x) \leq 1$ for every x. Then the series becomes a series of numbers, and the jth summand does not exceed $(3/4)^j$ in absolute value. Thus the series converges absolutely; therefore it converges. So it is clear that the displayed formula defines a function of x.

Step I: f is continuous. To see that f is continuous, pick an $\epsilon > 0$. Choose N so large that

$$\sum_{j=N+1}^{\infty} \left(\frac{3}{4}\right)^j < \frac{\epsilon}{4}$$

(we can of course do this because the series $\sum \left(\frac{3}{4}\right)^j$ converges). Now fix x. Observe that, since ψ is continuous and the graph of ψ is composed of segments of slope 1, we have

$$|\psi(s) - \psi(t)| \leq |s - t|$$

for all s and t. Moreover $|\psi(s) - \psi(t)| \leq 1$ for all s, t.

For $j = 1, 2, \ldots, N$, pick $\delta_j > 0$ so that, when $|t - x| < \delta_j$, then

$$\left|\psi\left(4^j t\right) - \psi\left(4^j x\right)\right| < \frac{\epsilon}{8}.$$

Let δ be the minimum of $\delta_1, \ldots \delta_N$.

Now, if $|t - x| < \delta$, then

$$
\begin{aligned}
|f(t) - f(x)| &= \left| \sum_{j=1}^{N} \left(\frac{3}{4}\right)^j \cdot \left(\psi(4^j t) - \psi(4^j x)\right) \right. \\
&\qquad \left. + \sum_{j=N+1}^{\infty} \left(\frac{3}{4}\right)^j \cdot \left(\psi(4^j t) - \psi(4^j x)\right) \right| \\
&\leq \sum_{j=1}^{N} \left(\frac{3}{4}\right)^j \left|\left(\psi(4^j t) - \psi(4^j x)\right)\right| \\
&\qquad + \sum_{j=N+1}^{\infty} \left(\frac{3}{4}\right)^j \left|\psi(4^j t) - \psi(4^j x)\right| \\
&\leq \sum_{j=1}^{N} \left(\frac{3}{4}\right)^j \cdot \frac{\epsilon}{8} + \sum_{j=N+1}^{\infty} \left(\frac{3}{4}\right)^j.
\end{aligned}
$$

Here we have used the choice of δ to estimate the summands in the first sum. The first sum is thus less than $\epsilon/2$ (just notice that $\sum_{j=1}^{\infty} (3/4)^j < 4$). The second sum is less than $\epsilon/2$ by the choice of N. Altogether then

$$
|f(t) - f(x)| < \epsilon
$$

whenever $|t - x| < \delta$. Therefore f is continuous, indeed uniformly so.

Step II: f is nowhere differentiable. Fix x. For $\ell = 1, 2, \ldots$ define $t_\ell = x \pm 4^{-\ell}/2$. We will say whether the sign is plus or minus in a moment (this will depend on the position of x relative to the integers). Then

$$
\begin{aligned}
\left| \frac{f(t_\ell) - f(x)}{t_\ell - x} \right| &= \left| \frac{1}{t_\ell - x} \left[\sum_{j=1}^{\ell} \left(\frac{3}{4}\right)^j \left(\psi(4^j t_\ell) - \psi(4^j x)\right) \right. \right. \\
&\qquad \left. \left. + \sum_{j=\ell+1}^{\infty} \left(\frac{3}{4}\right)^j \left(\psi(4^j t_\ell) - \psi(4^j x)\right) \right] \right|. \qquad (6.6.1)
\end{aligned}
$$

Notice that, when $j \geq \ell+1$, then $4^j t_\ell$ and $4^j x$ differ by an even integer. Since ψ has period 2, we find that each of the summands in the second sum is 0. Next we turn to the first sum.

We choose the sign—plus or minus—in the definition of t_ℓ so that there is no integer lying between $4^\ell t_\ell$ and $4^\ell x$. We can do this because the two numbers differ by $1/2$. But then the ℓth summand has magnitude

$$
(3/4)^\ell \cdot |4^\ell t_\ell - 4^\ell x| = 3^\ell |t_\ell - x|.
$$

On the other hand, the first $\ell - 1$ summands add up to not more than

$$\sum_{j=1}^{\ell-1} \left(\frac{3}{4}\right)^j \cdot |4^j t_\ell - 4^j x| = \sum_{j=1}^{\ell-1} 3^j \cdot 4^{-\ell}/2 \leq \frac{3^\ell - 1}{3 - 1} \cdot 4^{-\ell}/2 \leq 3^\ell \cdot 4^{-\ell-1}.$$

It follows that

$$
\begin{aligned}
\left| \frac{f(t_\ell) - f(x)}{t_\ell - x} \right| &= \frac{1}{|t_\ell - x|} \cdot \left| \sum_{j=1}^{\ell} \left(\frac{3}{4}\right)^j \left(\psi(4^j t_\ell) - \psi(4^j x) \right) \right| \\
&= \frac{1}{|t_\ell - x|} \cdot \left| \sum_{j=1}^{\ell-1} \left(\frac{3}{4}\right)^j \left(\psi(4^j t_\ell) - \psi(4^j x) \right) \right. \\
&\qquad \left. + \left(\frac{3}{4}\right)^\ell \left(\psi(4^\ell t_\ell) - \psi(4^\ell x) \right) \right| \\
&\geq \frac{1}{|t_\ell - x|} \cdot \left| \left(\frac{3}{4}\right)^\ell \psi(4^\ell t_\ell) - \left(\frac{3}{4}\right)^\ell \psi(4^\ell x) \right| \\
&\qquad - \frac{1}{|t_\ell - x|} \left| \sum_{j=1}^{\ell-1} \left(\frac{3}{4}\right)^j \left(\psi(4^j t_\ell) - \psi(4^j x) \right) \right| \\
&\geq 3^\ell - \frac{1}{(4^{-\ell}/2)} \cdot 3^\ell \cdot 4^{-\ell-1} \\
&\geq 3^{\ell-1}.
\end{aligned}
$$

Thus $t_\ell \to x$ but the Newton quotients blow up as $\ell \to \infty$. Therefore the limit

$$\lim_{t \to x} \frac{f(t) - f(x)}{t - x}$$

cannot exist. The function f is not differentiable at x.

$\square$

The proof of the last theorem was long, but the idea is simple: the function f is built by piling oscillations on top of oscillations. When the ℓth oscillation is added, it is made very small in size so that it does not cancel the previous oscillations. But it is made very steep so that it will cause the derivative to become large.

The practical meaning of Weierstrass's example is that we should realize that differentiability is a very strong and special property of functions. Most continuous functions are not differentiable at any point. When we are proving theorems about continuous functions, we should *not* think of them in terms of properties of differentiable functions.

Next we turn to the Chain Rule.

Theorem 6.7 *Let g be a differentiable function on an open interval I and let f be a differentiable function on an open interval that contains the range of g. Then $f \circ g$ is differentiable on the interval I and*

$$(f \circ g)'(x) = f'(g(x)) \cdot g'(x)$$

for each $x \in I$.

Proof: We use the notation Δt to stand for an increment in the variable t. Let us use the symbol $\mathcal{V}(r)$ to stand for any expression which tends to 0 as $\Delta r \to 0$. Fix $x \in I$. Set $r = g(x)$. By hypothesis,

$$\lim_{\Delta r \to 0} \frac{f(r + \Delta r) - f(r)}{\Delta r} = f'(r)$$

or

$$\frac{f(r + \Delta r) - f(r)}{\Delta r} - f'(r) = \mathcal{V}(r)$$

or

$$f(r + \Delta r) = f(r) + \Delta r \cdot f'(r) + \Delta r \cdot \mathcal{V}(r). \tag{6.7.1}$$

Notice that equation (6.7.1) is valid even when $\Delta r = 0$. Since Δr in equation (6.7.1) can be any small quantity, we set

$$\Delta r = \Delta x \cdot [g'(x) + \mathcal{V}(x)].$$

Substituting this expression into (6.7.1) and using the fact that $r = g(x)$ yields

$$f(g(x) + \Delta x[g'(x) + \mathcal{V}(x)]) =$$
$$f(r) + (\Delta x \cdot [g'(x) + \mathcal{V}(x)]) \cdot f'(r) + (\Delta x \cdot [g'(x) + \mathcal{V}(x)]) \cdot \mathcal{V}(r)$$
$$= f(g(x)) + \Delta x \cdot f'(g(x)) \cdot g'(x) + \Delta x \cdot \mathcal{V}(x). \tag{6.7.2}$$

Just as we derived (6.7.1), we may also obtain

$$\begin{aligned} g(x + \Delta x) &= g(x) + \Delta x \cdot g'(x) + \Delta x \cdot \mathcal{V}(x) \\ &= g(x) + \Delta x[g'(x) + \mathcal{V}(x)]. \end{aligned}$$

We may substitute this equality into the left side of (6.7.2) to obtain

$$f(g(x + \Delta x)) = f(g(x)) + \Delta x \cdot f'(g(x)) \cdot g'(x) + \Delta x \cdot \mathcal{V}(x).$$

With some algebra this can be rewritten as

$$\frac{f(g(x + \Delta x)) - f(g(x))}{\Delta x} - f'(g(x)) \cdot g'(x) = \mathcal{V}(x).$$

But this just says that

$$\lim_{\Delta x \to 0} \frac{(f \circ g)(x + \Delta x) - (f \circ g)(x)}{\Delta x} = f'(g(x)) \cdot g'(x).$$

That is, $(f \circ g)'(x)$ exists and equals $f'(g(x)) \cdot g'(x)$, as desired. □

EXAMPLE 6.8 The derivative of

$$f(x) = \sin(x^3 - x^2)$$

is

$$f'(x) = [\cos(x^3 - x^2)] \cdot (3x^2 - 2x).$$ □

Exercises

1. For which positive integers k is it true that if $f^k = f \cdot f \cdots f$ is differentiable at x then f is differentiable at x?

2. Let f be a function that has domain an interval I and takes values in the complex numbers. Then we may write $f(x) = u(x) + iv(x)$ with u and v each being real-valued functions. We say that f is differentiable at a point $x \in I$ if both u and v are. Formulate an alternative definition of differentiability of f at a point x which makes no reference to u and v (but instead defines the derivative directly in terms of f) and prove that your new definition is equivalent to the definition in terms of u and v.

3. Let $f(x)$ equal 0 if x is irrational; let $f(x)$ equal $1/q$ if x is a rational number that can be expressed in lowest terms as p/q. Is f differentiable at any x?

4. Assume that f is a continuous function on $(-1, 1)$ and that f is differentiable on $(-1, 0) \cup (0, 1)$. If the limit $\lim_{x \to 0} f'(x)$ exists then is f differentiable at $x = 0$?

5. Formulate notions of "left differentiable" and "right differentiable" for functions defined on suitable half-open intervals. Also formulate definitions of "left continuous" and "right continuous." If you have done things correctly, then you should be able to prove that a left differentiable (right differentiable) function is left continuous (right continuous).

6. Define
$$f(x) = \begin{cases} x^{3/2} \cdot \sin(1/x) & \text{if} \quad x \neq 0 \\ 0 & \text{if} \quad x = 0. \end{cases}$$

Prove that f is differentiable at every point, but that the derivative function f' is discontinuous at 0.

7. Refer to Exercise **6**. Is the discontinuity at the origin of the first kind or the second kind?

8. Prove part (a) of Theorem 6.4.

9. Verify the properties of the derivative presented in Theorem 6.4 in the new context of complex-valued functions.

*   **10.** Let $E \subseteq \mathbb{R}$ be a closed set. Fix a nonnegative integer k. Show that there is a function f in $C^k(\mathbb{R})$ (that is, a k-times continuously differentiable function) such that $E = \{x : f(x) = 0\}$.

*   **11.** Prove that the nowhere differentiable function constructed in Theorem 6.6 is in Lip_α for all $\alpha < 1$.

*   **12.** Prove that the Weierstrass nowhere differentiable function f constructed in Theorem 6.6 satisfies

    $$\frac{|f(x+h) + f(x-h) - 2f(x)|}{|h|} \leq C|h|$$

    for all nonzero h but f is *not* Lipschitz-1.

    13. Fill in the details for this alternative proof of the product rule for the derivative:

    (a) Prove that $(f^2)' = 2f \cdot f'$.

    (b) Apply the result of part **(a)** to the function $f + g$.

    (c) Use the result of part **(a)** to cancel terms in the formula from part **(b)** to obtain the product rule.

6.2   The Mean Value Theorem and Applications

We begin this section with some remarks about local maxima and minima of functions.

Definition 6.9 Let f be a function with domain (a, b). A point $C \in (a, b)$ is called a *local maximum* for f (we also say that f has a local maximum at C) if there is a $\delta > 0$ such that $f(t) \leq f(C)$ for all $t \in (C - \delta, C + \delta)$. A point $c \in (a, b)$ is called a *local minimum* for f (we also say that f has a local minimum at c) if there is a $\delta > 0$ such that $f(t) \geq f(c)$ for all $t \in (c - \delta, c + \delta)$. See Figure 6.4.

Local minima (plural of minimum) and local maxima (plural of maximum) are referred to collectively as *local extrema.*

Proposition 6.10 (Fermat) *If f is a function with domain (a, b), if f has a local extremum at $x \in (a, b)$, and if f is differentiable at x, then $f'(x) = 0$.*

Proof: Suppose that f has a local minimum at x. Then there is a $\delta > 0$ such that if $x - \delta < t < x$ then $f(t) \geq f(x)$. Then

$$\frac{f(t) - f(x)}{t - x} \leq 0 \, .$$

Letting $t \to x$, it follows that $f'(x) \leq 0$. Similarly, if $x < t < x + \delta$ for suitable δ, then

$$\frac{f(t) - f(x)}{t - x} \geq 0 \, .$$
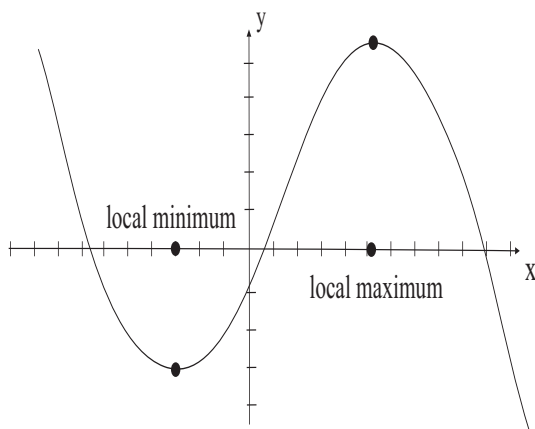
Figure 6.4: Some extrema.

It follows that $f'(x) \geq 0$. We must conclude that $f'(x) = 0$.

A similar argument applies if f has a local maximum at x. The proof is therefore complete. □

EXAMPLE 6.11 Consider the function $f(x) = \sin x$ on the interval $[\pi/3, 11\pi/3]$. Surely $f'(x) = \cos x$, and we see that f' vanishes at the points $\pi/2$, $3\pi/2$, $5\pi/2$, $7\pi/2$ of the interval. By Fermat's theorem, these are candidates to be local maxima or minima of f. And, indeed, $\pi/2$, $5\pi/2$ are local maxima and $3\pi/2$, $7\pi/2$ are local minima. □

Before going on to mean value theorems, we provide a striking application of the proposition:

Theorem 6.12 (Darboux's Theorem) *Let f be a differentiable function on an open interval I. Pick points $s < t$ in I and suppose that $f'(s) < \rho < f'(t)$. Then there is a point u between s and t such that $f'(u) = \rho$.*

Proof: Consider the function $g(x) = f(x) - \rho x$. Then $g'(s) < 0$ and $g'(t) > 0$. Assume for simplicity that $s < t$. The sign of the derivative at s shows that $g(\widehat{s}) < g(s)$ for $\widehat{s}$ greater than s and near s. The sign of the derivative at t implies that $g(\widehat{t}) < g(t)$ for $\widehat{t}$ less than t and near t. Thus the minimum of the continuous function g on the compact interval $[s, t]$ must occur at some point u in the interior (s, t). The preceding proposition guarantees that $g'(u) = 0$, or $f'(u) = \rho$ as claimed. □

EXAMPLE 6.13 If f' were a continuous function then the theorem would just be a special instance of the Intermediate Value Property of continuous functions
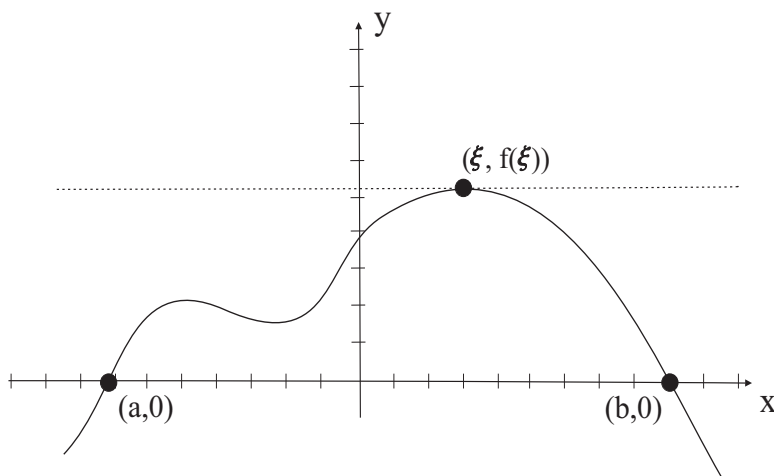
Figure 6.5: Rolle's theorem.

(see Corollary 5.31). But derivatives need not be continuous, as the example

$$f(x) = \begin{cases} x^{3/2} \cdot \sin(1/x) & \text{if} \quad x \neq 0 \\ 0 & \text{if} \quad x = 0 \end{cases}$$

illustrates. Check for yourself that $f'(0)$ exists and vanishes but $\lim_{x \to 0} f'(x)$ does not exist. This example illustrates the significance of the theorem.   □

Since the theorem says that f' will always satisfy the Intermediate Value Property (even when it is not continuous), its discontinuities cannot be of the first kind. In other words:

Proposition 6.14 *If f is a differentiable function on an open interval I then the discontinuities of f' are all of the second kind.*

Next we turn to the simplest form of the Mean Value Theorem.

Theorem 6.15 (Rolle's Theorem) *Let f be a continuous function on the closed interval $[a, b]$ which is differentiable on (a, b). If $f(a) = f(b) = 0$ then there is a point $\xi \in (a, b)$ such that $f'(\xi) = 0$. See Figure 6.5.*

Proof: If f is a constant function then any point ξ in the interval will do. So assume that f is nonconstant.

Theorem 5.23 guarantees that f will have both a maximum and a minimum in $[a, b]$. If one of these occurs in (a, b) then Proposition 6.10 guarantees that f' will vanish at that point and we are done. If both occur at the endpoints then all the values of f lie between 0 and 0. In other words f is constant, contradicting our assumption.   □
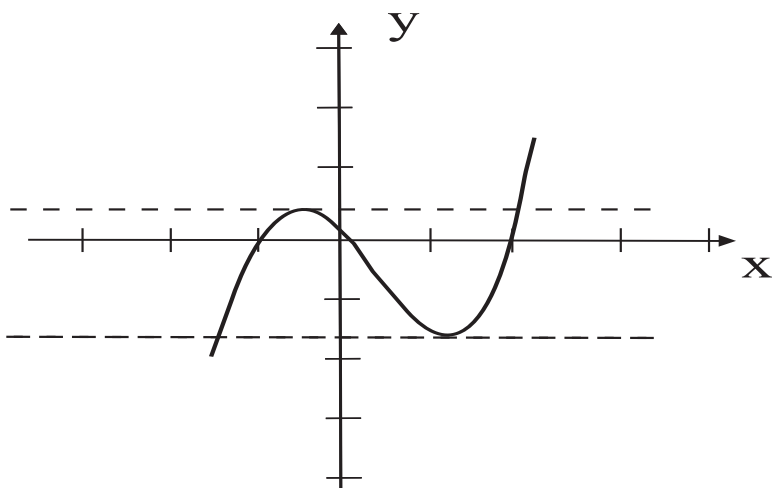
Figure 6.6: An example of Rolle's theorem.

EXAMPLE 6.16 Of course the point ξ in Rolle's theorem need not be unique. If $f(x) = x^3 - x^2 - 2x$ on the interval $[-1, 2]$ then $f(-1) = f(2) = 0$ and $f'(x) = 3x^2 - 2x - 2$ vanishes at *two* points (namely, $(2+\sqrt{28})/6$ and $(2-\sqrt{28})/6$) of the interval $(-1, 2)$. Refer to Figure 6.6. □

If you rotate the graph of a function satisfying the hypotheses of Rolle's theorem, the result suggests that, for any continuous function f on an interval $[a, b]$, differentiable on (a, b), we should be able to relate the slope of the chord connecting $(a, f(a))$ and $(b, f(b))$ with the value of f' at some interior point. That is the content of the standard Mean Value Theorem:

Theorem 6.17 (The Mean Value Theorem) *Let f be a continuous function on the closed interval $[a, b]$ that is differentiable on (a, b). There exists a point $\xi \in (a, b)$ such that*

$$\frac{f(b) - f(a)}{b - a} = f'(\xi).$$

See Figure 6.7.

Proof: Our scheme is to implement the remarks preceding the theorem: we "rotate" the picture to reduce to the case of Rolle's theorem. More precisely, define

$$g(x) = f(x) - \left[f(a) + \frac{f(b) - f(a)}{b - a} \cdot (x - a) \right] \quad \text{if} \quad x \in [a, b].$$

By direct verification, g is continuous on $[a, b]$ and differentiable on (a, b) (after all, g is obtained from f by elementary arithmetic operations). Also $g(a) =$
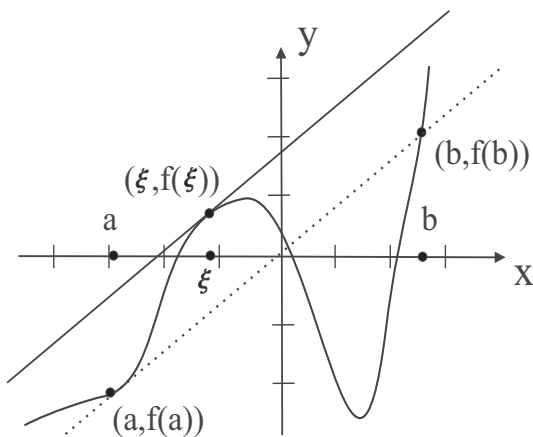
Figure 6.7: The Mean Value Theorem.

$g(b) = 0$. Thus we may apply Rolle's theorem to g and we find that there is a $\xi \in (a, b)$ such that $g'(\xi) = 0$. Remembering that x is the variable, we differentiate the formula for g to find that

$$0 = g'(\xi) \quad = \quad \left[f'(x) - \frac{f(b) - f(a)}{b - a} \right]\Bigg|_{x=\xi}$$

$$= \quad \left[f'(\xi) - \frac{f(b) - f(a)}{b - a} \right].$$

As a result,

$$f'(\xi) = \frac{f(b) - f(a)}{b - a}. \qquad \qquad \square$$

Corollary 6.18 *If f is a differentiable function on the open interval I and if $f'(x) = 0$ for all $x \in I$ then f is a constant function.*

Proof: If s and t are any two elements of I then the theorem tells us that

$$f(s) - f(t) = f'(\xi) \cdot (s - t)$$

for some ξ between s and t. But, by hypothesis, $f'(\xi) = 0$. We conclude that $f(s) = f(t)$. But, since s and t were chosen arbitrarily, we must conclude that f is constant. $\qquad \square$

Corollary 6.19 *If f is differentiable on an open interval I and $f'(x) \geq 0$ for all $x \in I$, then f is increasing on I; that is, if $s < t$ are elements of I, then $f(s) \leq f(t)$.*

If f is differentiable on an open interval I and $f'(x) \leq 0$ for all $x \in I$, then f is decreasing on I; that is, if $s < t$ are elements of I, then $f(s) \geq f(t)$.

Proof: Similar to the preceding corollary.   □

EXAMPLE 6.20 Let us verify that, if f is a differentiable function on $\mathbb{R}$, and if $|f'(x)| \leq 1$ for all x, then $|f(s) - f(t)| \leq |s - t|$ for all real s and t.

In fact, for $s \neq t$ there is a ξ between s and t such that

$$\frac{f(s) - f(t)}{s - t} = f'(\xi) \,.$$

But $|f'(\xi)| \leq 1$ by hypothesis hence

$$\left| \frac{f(s) - f(t)}{s - t} \right| = |f'(\xi)| \leq 1$$

or

$$|f(s) - f(t)| \leq |s - t| \,. \qquad\qquad □$$

EXAMPLE 6.21 Let us verify that

$$\lim_{x \to +\infty} \left(\sqrt{x + 5} - \sqrt{x} \right) = 0 \,.$$

Here the limit operation means that, for any $\epsilon > 0$, there is an $N > 0$ such that $x > N$ implies that the expression in parentheses has absolute value less than ϵ.

Define $f(x) = \sqrt{x}$ for $x > 0$. Then the expression in parentheses is just $f(x + 5) - f(x)$. By the Mean Value Theorem this equals

$$f'(\xi) \cdot 5$$

for some $x < \xi < x + 5$. But this last expression is

$$\frac{1}{2} \cdot \xi^{-1/2} \cdot 5 \,.$$

By the bounds on ξ, this is

$$\leq \frac{5}{2} x^{-1/2} \,.$$

Clearly, as $x \to +\infty$, this expression tends to zero.   □

A powerful tool in analysis is a generalization of the usual Mean Value Theorem that is due to A. L. Cauchy (1789–1857):

Theorem 6.22 (Cauchy's Mean Value Theorem) *Let f and g be continuous functions on the interval $[a, b]$ which are both differentiable on the interval (a, b), $a < b$. Assume that $g' \neq 0$ on the interval and that $g(a) \neq g(b)$. Then there is a point $\xi \in (a, b)$ such that*

$$\frac{f(b) - f(a)}{g(b) - g(a)} = \frac{f'(\xi)}{g'(\xi)} \,.$$

Proof: Apply the usual Mean Value Theorem to the function

$$h(x) = g(x) \cdot \{f(b) - f(a)\} - f(x) \cdot \{g(b) - g(a)\} \,. \qquad \square$$

Clearly the usual Mean Value Theorem (Theorem 6.17) is obtained from Cauchy's by taking $g(x)$ to be the function x. We conclude this section by illustrating a typical application of the result.

EXAMPLE 6.23 Let f be a differentiable function on an interval I such that f' is differentiable at a point $x \in I$. Then

$$\lim_{h \to 0^+} \frac{f(x+h) + f(x-h) - 2f(x)}{h^2} = (f')'(x) \equiv f''(x) \,.$$

To see this, fix x and define $\mathcal{F}(h) = f(x+h) + f(x-h) - 2f(x)$ and $\mathcal{G}(h) = h^2$. Then

$$\frac{f(x+h) + f(x-h) - 2f(x)}{h^2} = \frac{\mathcal{F}(h) - \mathcal{F}(0)}{\mathcal{G}(h) - \mathcal{G}(0)} \,.$$

According to Cauchy's Mean Value Theorem, there is a ξ between 0 and h such that the last line equals

$$\frac{\mathcal{F}'(\xi)}{\mathcal{G}'(\xi)} \,.$$

Writing this last expression out gives

$$\frac{f'(x+\xi) - f'(x-\xi)}{2\xi} = \frac{1}{2} \cdot \frac{f'(x+\xi) - f'(x)}{\xi}$$
$$+ \frac{1}{2} \cdot \frac{f'(x-\xi) - f'(x)}{-\xi} \,,$$

and the last line tends as $h \to 0$, by the definition of the derivative, to the quantity $(f')'(x)$. $\qquad \square$

It is a fact that the standard proof of l'Hôpital's Rule (Guillaume François Antoine de l'Hôpital, Marquis de St.-Mesme, 1661–1704) is obtained by way of Cauchy's Mean Value Theorem. This line of reasoning is explored in the next section.

Exercises

1. Let f be a function that is continuous on $[0, \infty)$ and differentiable on $(0, \infty)$. If $f(0) = 0$ and $|f'(x)| \le |f(x)|$ for all $x > 0$ then prove that $|f(x)| \le e^x$ for all x. [This result is often called Gronwall's inequality.]

2. Let f be a continuous function on $[a, b]$ that is differentiable on (a, b). Assume that $f(a) = m$ and that $|f'(x)| \le K$ for all $x \in (a, b)$. What bound can you then put on the magnitude of $f(b)$?

3. Let f be a differentiable function on an open interval I and assume that f has no local minima nor local maxima on I. Prove that f is either increasing or decreasing on I.

4. Let $0 < \alpha \leq 1$. Prove that there is a constant $C_\alpha > 0$ such that, for $0 < x < 1$, it holds that
$$|\ln x| \leq C_\alpha \cdot x^{-\alpha}.$$
Prove that the constant cannot be taken to be independent of α.

5. Let f be a function that is twice continuously differentiable on $[0, \infty)$ and assume that $f''(x) \geq c > 0$ for all x. Prove that f is not bounded from above.

6. Let f be differentiable on an interval I and $f'(x) > 0$ for all $x \in I$. Does it follow that $(f^2)' > 0$ for all $x \in I$? What additional hypothesis on f will make the conclusion true?

7. Answer Exercise **6** with the exponent 2 replaced by any positive integer exponent.

8. Use the Mean Value Theorem to say something about the behavior at $+\infty$ of the function $g(x) = \sqrt[3]{x^4 + 1} - x^{4/3}$.

9. Use the Mean Value Theorem to say something about the behavior at $+\infty$ of the function $f(x) = \sqrt{x+1} - \sqrt{x}$.

10. Refer to Exercise **9**. What can you say about the asymptotics at $+\infty$ of $\sqrt{x+1}/\sqrt{x}$?

11. Supply the details of the proof of Theorem 6.22.

12. Give an example of a function f for which the limit in Example 6.23 exists at some x but for which f is not twice differentiable at x.

6.3   More on the Theory of Differentiation

l'Hôpital's Rule (actually due to his teacher J. Bernoulli (1667–1748)) is a useful device for calculating limits, and a nice application of the Cauchy Mean Value Theorem. Here we present a special case of the theorem.

Theorem 6.24 *Suppose that f and g are differentiable functions on an open interval I and that $p \in I$. If $\lim_{x \to p} f(x) = \lim_{x \to p} g(x) = 0$ and if*
$$\lim_{x \to p} \frac{f'(x)}{g'(x)} \tag{6.24.1}$$
exists and equals a real number ℓ then
$$\lim_{x \to p} \frac{f(x)}{g(x)} = \ell.$$

Proof: Fix a real number $a > \ell$. By (6.24.1) there is a number $q > p$ such that, if $p < x < q$, then

$$\frac{f'(x)}{g'(x)} < a\,. \tag{6.24.2}$$

But now, if $p < s < t < q$, then

$$\frac{f(t) - f(s)}{g(t) - g(s)} = \frac{f'(x)}{g'(x)}$$

for some $s < x < t$ (by Cauchy's Mean Value Theorem). It follows then from (6.24.2) that

$$\frac{f(t) - f(s)}{g(t) - g(s)} < a\,.$$

Now let $s \to p$ and invoke the hypothesis about the zero limit of f and g at p to conclude that

$$\frac{f(t)}{g(t)} \le a$$

when $p < t < q$. Since a is an arbitrary number to the right of ℓ we conclude that

$$\limsup_{t \to p^+} \frac{f(t)}{g(t)} \le \ell\,.$$

Similar arguments show that

$$\liminf_{t \to p^+} \frac{f(t)}{g(t)} \ge \ell\,;$$

$$\limsup_{t \to p^-} \frac{f(t)}{g(t)} \le \ell\,;$$

$$\liminf_{t \to p^-} \frac{f(t)}{g(t)} \ge \ell\,.$$

We conclude that the desired limit exists and equals ℓ. $\square$

A closely related result, with a similar proof, is this:

Theorem 6.25 *Suppose that f and g are differentiable functions on an open interval I and that $p \in I$. If $\lim_{x \to p} f(x) = \lim_{x \to p} g(x) = \pm\infty$ and if*

$$\lim_{x \to p} \frac{f'(x)}{g'(x)} \tag{6.25.1}$$

exists and equals a real number ℓ then

$$\lim_{x \to p} \frac{f(x)}{g(x)} = \ell\,.$$

EXAMPLE 6.26 Let
$$f(x) = \left| \ln |x| \right|^{(x^2)}.$$

We wish to determine $\lim_{x \to 0} f(x)$. To do so, we define

$$F(x) = \ln f(x) = x^2 \ln \left| \ln |x| \right| = \frac{\ln \left| \ln |x| \right|}{1/x^2}.$$

Notice that both the numerator and the denominator tend to $\pm \infty$ as $x \to 0$. So the hypotheses of l'Hôpital's rule are satisfied and the limit is

$$\lim_{x \to 0} \frac{\ln \left| \ln |x| \right|}{1/x^2} = \lim_{x \to 0} \frac{(\ln \left| \ln |x| \right|)'}{(1/x^2)'} = \lim_{x \to 0} \frac{1/[x \ln |x|]}{-2/x^3} = \lim_{x \to 0} \frac{-x^2}{2 \ln |x|} = 0.$$

Since $\lim_{x \to 0} F(x) = 0$ we may calculate that the original limit has value $\lim_{x \to 0} f(x) = 1$. □

Proposition 6.27 Let f be an invertible function on an interval (a, b) with nonzero derivative at a point $x \in (a, b)$. Let $X = f(x)$. Then $(f^{-1})'(X)$ exists and equals $1/f'(x)$.

Proof: Observe that, for $T \neq X$,

$$\frac{f^{-1}(T) - f^{-1}(X)}{T - X} = \frac{1}{\frac{f(t) - f(x)}{t - x}}, \tag{6.27.1}$$

where $T = f(t)$. Since $f'(x) \neq 0$, the difference quotients for f in the denominator are bounded from zero hence the limit of the formula in (6.27.1) exists. This proves that f^{-1} is differentiable at X and that the derivative at that point equals $1/f'(x)$. □

EXAMPLE 6.28 We know that the function $f(x) = x^k$, k a positive integer, is one-to-one and differentiable on the interval $(0, 1)$. Moreover the derivative $k \cdot x^{k-1}$ never vanishes on that interval. Therefore the proposition applies and we find for $X \in (0, 1) = f((0, 1))$ that

$$\left(f^{-1} \right)'(X) = \frac{1}{f'(x)} = \frac{1}{f'(X^{1/k})}$$

$$= \frac{1}{k \cdot X^{1-1/k}} = \frac{1}{k} \cdot X^{1/k-1}.$$

In other words,

$$\left(X^{1/k} \right)' = \frac{1}{k} X^{1/k-1}. \qquad \square$$

We conclude this section by saying a few words about higher derivatives. If f is a differentiable function on an open interval I then we may ask whether the function f' is differentiable. If it is, then we denote its derivative by

$$f'' \quad \text{or} \quad f^{(2)} \quad \text{or} \quad \frac{d^2}{dx^2}f \quad \text{or} \quad \frac{d^2 f}{dx^2},$$

and call it the second derivative of f. Likewise the derivative of the $(k-1)$th derivative, if it exists, is called the kth derivative and is denoted

$$f''^{\cdots'} \quad \text{or} \quad f^{(k)} \quad \text{or} \quad \frac{d^k}{dx^k}f \quad \text{or} \quad \frac{d^k f}{dx^k}.$$

Observe that we cannot even consider whether $f^{(k)}$ exists at a point unless $f^{(k-1)}$ exists in a *neighborhood* of that point.

If f is k times differentiable on an open interval I and if each of the derivatives $f^{(1)}, f^{(2)}, \ldots, f^{(k)}$ is continuous on I then we say that the function f is k *times continuously differentiable* on I. We write $f \in C^k(I)$. Obviously there is some redundancy in this definition since the continuity of $f^{(j-1)}$ follows from the existence of $f^{(j)}$. Thus only the continuity of the last derivative $f^{(k)}$ need be checked. Continuously differentiable functions are useful tools in analysis. We denote the class of k times continuously differentiable functions on I by $C^k(I)$.

EXAMPLE 6.29 For $k = 1, 2, \ldots$ the function

$$f_k(x) = \begin{cases} x^{k+1} & \text{if} \quad x \geq 0 \\ -x^{k+1} & \text{if} \quad x < 0 \end{cases}$$

will be k times continuously differentiable on $\mathbb{R}$ but will fail to be $k + 1$ times differentiable at $x = 0$. More dramatically, an analysis similar to the one we used on the Weierstrass nowhere differentiable function shows that the function

$$g_k(x) = \sum_{j=1}^{\infty} \frac{3^j}{4^{j+jk}} \sin(4^j x)$$

is k times continuously differentiable on $\mathbb{R}$ but will not be $k + 1$ times differentiable at any point (this function, with $k = 0$, was Weierstrass's original example). $\square$

A more refined notion of smoothness/continuity of functions is that of Hölder continuity or Lipschitz continuity (see Section 5.3). If f is a function on an open interval I and if $0 < \alpha \leq 1$ then we say that f satisfies a *Lipschitz condition* of order α on I if there is a constant M such that for all $s, t \in I$ we have

$$|f(s) - f(t)| \leq M \cdot |s - t|^{\alpha}.$$

Such a function is said to be of class $\text{Lip}_{\alpha}(I)$. Clearly a function of class Lip_{α} is uniformly continuous on I. For, if $\epsilon > 0$, then we may take $\delta = (\epsilon/M)^{1/\alpha}$: it follows that, for $|s - t| < \delta$, we have

$$|f(s) - f(t)| \leq M \cdot |s - t|^{\alpha} < M \cdot \epsilon/M = \epsilon.$$

Interestingly, when $\alpha > 1$ the class Lip_α contains only constant functions. For in this instance the inequality

$$|f(s) - f(t)| \leq M \cdot |s - t|^\alpha$$

leads to

$$\left| \frac{f(s) - f(t)}{s - t} \right| \leq M \cdot |s - t|^{\alpha - 1}.$$

Because $\alpha - 1 > 0$, letting $s \to t$ yields that $f'(t)$ exists for every $t \in I$ and equals 0. It follows from Corollary 6.18 of the last section that f is constant on I.

Instead of trying to extend the definition of $\text{Lip}_\alpha(I)$ to $\alpha > 1$ it is customary to define classes of functions $C^{k,\alpha}$, for $k = 0, 1, \ldots$ and $0 < \alpha \leq 1$, by the condition that f be of class C^k on I and that $f^{(k)}$ be an element of $\text{Lip}_\alpha(I)$. We leave it as an exercise for you to verify that $C^{k,\alpha} \subseteq C^{\ell,\beta}$ if either $k > \ell$ or both $k = \ell$ and $\alpha \geq \beta$.

In more advanced studies in analysis, it is appropriate to replace $\text{Lip}_1(I)$, and more generally $C^{k,1}$, with another space (invented by Antoni Zygmund, 1900–1992) defined in a more subtle fashion using second differences as in Example 6.23. These matters exceed the scope of this book, but we shall make a few remarks about them in the exercises.

Exercises

1. Suppose that f is a C^2 function on $\mathbb{R}$ and that $|f''(x)| \leq C$ for all x. Prove that
$$\left| \frac{f(x + h) + f(x - h) - 2f(x)}{h^2} \right| \leq C.$$

2. Fix a positive integer k. Give an example of two functions f and g neither of which is in C^k but such that $f \cdot g \in C^k$.

3. Fix a positive integer ℓ and define $f(x) = |x|^\ell$. In which class C^k does f lie? In which class $C^{k,\alpha}$ does it lie?

4. In the text we give sufficient conditions for the inclusion $C^{k,\alpha} \subseteq C^{\ell,\beta}$. Show that the inclusion is strict if either $k > \ell$ or $k = \ell$ and $\alpha > \beta$.

5. Suppose that f is a differentiable function on an interval I and that $f'(x)$ is never zero. Prove that f is invertible. Then prove that f^{-1} is differentiable. Finally, use the Chain Rule on the identity $f\left(f^{-1}\right) = x$ to derive a formula for $\left(f^{-1}\right)'$.

6. Suppose that a function f on the interval $(0, 1)$ has left derivative equal to zero at every point. What conclusion can you draw?

7. We know that the first derivative can be characterized by the Newton quotient. Find an analogous characterization of second derivatives. What about third derivatives?

8. Use l'Hôpital's Rule to analyze the limit

$$\lim_{x \to +\infty} x^{1/x}.$$

* 9. We know (see Section 8.4) that a continuous function on the interval $[0, 1]$ can be uniformly approximated by polynomials. But if the function f is continuously differentiable on $[0, 1]$ then we can actually say something about the *rate* of approximation. That is, if $\epsilon > 0$ then f can be approximated uniformly within ϵ by a polynomial of degree not greater than $N = N(\epsilon)$. Calculate $N(\epsilon)$.

* 10. In which class $C^{k,\alpha}$ is the function $x \cdot \ln |x|$ on the interval $[-1/2, 1/2]$? How about the function $x/\ln |x|$?

* 11. Give an example of a function on $\mathbb{R}$ such that

$$\left| \frac{f(x+h) + f(x-h) - 2f(x)}{h} \right| \le C$$

for all x and all $h \neq 0$ but f is not in $\mathrm{Lip}_1(\mathbb{R})$. (**Hint:** See Exercise **3**.)

Chapter 7

The Integral

7.1 Partitions and the Concept of Integral

We learn in calculus that it is often useful to think of an integral as representing area. However, this is but one of many important applications of integration theory. The integral is a generalization of the summation process. That is the point of view that we shall take in the present chapter.

Definition 7.1 Let $[a, b]$ be a closed interval in $\mathbb{R}$. A finite, ordered set of points $\mathcal{P} = \{x_0, x_1, x_2, \ldots, x_{k-1}, x_k\}$ such that

$$a = x_0 \leq x_1 \leq x_2 \leq \cdots \leq x_{k-1} \leq x_k = b$$

is called a *partition* of $[a, b]$. Refer to Figure 7.1.

If $\mathcal{P}$ is a partition of $[a, b]$, then we let I_j denote the interval $[x_{j-1}, x_j]$, $j = 1, 2, \ldots, k$. The symbol Δ_j denotes the *length* of I_j. The *mesh* of $\mathcal{P}$, denoted by $m(\mathcal{P})$, is defined to be $\max_j \Delta_j$.

The points of a partition need not be equally spaced, nor must they be distinct from each other.

EXAMPLE 7.2 The set $\mathcal{P} = \{0, 1, 1, 9/8, 2, 5, 21/4, 23/4, 6\}$ is a partition of the interval $[0, 6]$ with mesh 3 (because $I_5 = [2, 5]$, with length 3, is the longest interval in the partition). See Figure 7.2. □
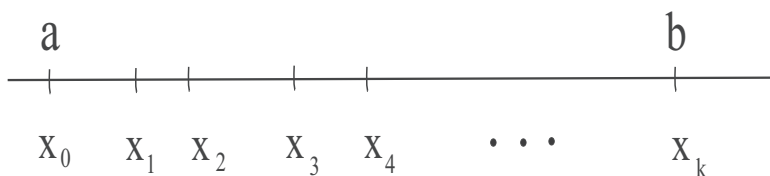
Figure 7.1: A partition.

Figure 7.2: The partition in Example 7.2.

Definition 7.3 Let $[a, b]$ be an interval and let f be a function with domain $[a, b]$. If $\mathcal{P} = \{x_0, x_1, x_2, \ldots, x_{k-1}, x_k\}$ is a partition of $[a, b]$ and if, for each j, s_j is an element of I_j, then the corresponding *Riemann sum* is defined to be

$$\mathcal{R}(f, \mathcal{P}) = \sum_{j=1}^{k} f(s_j) \Delta_j.$$

EXAMPLE 7.4 Let $f(x) = x^2 - x$ and $[a, b] = [1, 4]$. Define the partition $\mathcal{P} = \{1, 3/2, 2, 7/3, 4\}$ of this interval. Then a Riemann sum for this f and $\mathcal{P}$ is

$$\mathcal{R}(f, \mathcal{P}) = \left(1^2 - 1\right) \cdot \frac{1}{2} + \left((7/4)^2 - (7/4)\right) \cdot \frac{1}{2}$$
$$+ \left((7/3)^2 - (7/3)\right) \cdot \frac{1}{3} + \left(3^2 - 3\right) \cdot \frac{5}{3}$$
$$= \frac{10103}{864}.\qquad\qquad\qquad\square$$

Notice that we have complete latitude in choosing each point s_j from the corresponding interval I_j. While at first confusing, we will find this freedom to be a powerful tool when proving results about the integral.

The first main step in the theory of the Riemann integral is to determine a method for "calculating the limit of the Riemann sums" of a function as the mesh of the partitions tends to zero. There are in fact several methods for doing so. We have chosen the simplest one.

Definition 7.5 Let $[a, b]$ be an interval and f a function with domain $[a, b]$. We say that *the Riemann sums of f tend to a limit ℓ as $m(\mathcal{P})$ tends to 0* if, for any $\epsilon > 0$, there is a $\delta > 0$ such that, if $\mathcal{P}$ is any partition of $[a, b]$ with $m(\mathcal{P}) < \delta$, then $|\mathcal{R}(f, \mathcal{P}) - \ell| < \epsilon$ for every choice of $s_j \in I_j$.

It will turn out to be critical for the success of this definition that we require that *every* partition of mesh smaller than δ satisfy the conclusion of the definition. The theory does not work effectively if for every $\epsilon > 0$ there is a $\delta > 0$ and *some* partition $\mathcal{P}$ of mesh less than δ which satisfies the conclusion of the definition.

Definition 7.6 A function f on a closed interval $[a, b]$ is said to be *Riemann integrable* on $[a, b]$ if the Riemann sums of $\mathcal{R}(f, \mathcal{P})$ tend to a finite limit ℓ as $m(\mathcal{P})$ tends to zero.

The value ℓ of the limit, when it exists, is called the *Riemann integral* of f over $[a, b]$ and is denoted by

$$\int_a^b f(x)\,dx\,.$$

Remark 7.7 We mention now a useful fact that will be formalized in later sections. Suppose that f is Riemann integrable on $[a, b]$ with the value of the integral being ℓ. Let $\epsilon > 0$. Then, as stated in the definition (with $\epsilon/2$ replacing ϵ), there is a $\delta > 0$ such that, if $\mathcal{Q}$ is a partition of $[a, b]$ of mesh smaller than δ, then $|\mathcal{R}(f, \mathcal{Q}) - \ell| < \epsilon/2$. It follows that, if $\mathcal{P}$ and $\mathcal{P}'$ are partitions of $[a, b]$ of mesh smaller than δ, then

$$|\mathcal{R}(f, \mathcal{P}) - \mathcal{R}(f, \mathcal{P}')| \le |\mathcal{R}(f, \mathcal{P}) - \ell| + |\ell - \mathcal{R}(f, \mathcal{P}')| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon\,.$$

This is like a Cauchy condition.

Note, however, that we may choose $\mathcal{P}'$ to equal the partition $\mathcal{P}$. Also we may for each j choose the point s_j, where f is evaluated for the Riemann sum over $\mathcal{P}$, to be a point where f very nearly assumes its supremum on I_j. Likewise we may for each j choose the point s_j', where f is evaluated for the Riemann sum over $\mathcal{P}'$, to be a point where f very nearly assumes its infimum on I_j. It easily follows that, when the mesh of $\mathcal{P}$ is less than δ, then

$$\sum_j \left(\sup_{I_j} f - \inf_{I_j} f \right) \Delta_j \le \epsilon\,. \tag{7.7.1}$$

This consequence of integrability will prove useful to us in some of the discussions in this and the next section. In the exercises we shall consider in detail the assertion that integrability implies (7.7.1) and the converse as well.

Definition 7.8 If $\mathcal{P}, \mathcal{P}'$ are partitions of $[a, b]$ then their *common refinement* is the union of all the points of $\mathcal{P}$ and $\mathcal{P}'$. See Figure 7.3.

We record now a technical lemma that will be used in several of the proofs that follow:

Lemma 7.9 *Let f be a function with domain the closed interval $[a, b]$. The Riemann integral*

$$\int_a^b f(x)\,dx$$

exists if and only if, for every $\epsilon > 0$, there is a $\delta > 0$ such that, if $\mathcal{P}$ and $\mathcal{P}'$ are partitions of $[a, b]$ with $m(\mathcal{P}) < \delta$ and $m(\mathcal{P}') < \delta$, then their common refinement $\mathcal{Q}$ has the property that

$$|\mathcal{R}(f, \mathcal{P}) - \mathcal{R}(f, \mathcal{Q})| < \epsilon$$

and (7.9.1)

$$|\mathcal{R}(f, \mathcal{P}') - \mathcal{R}(f, \mathcal{Q})| < \epsilon\,.$$
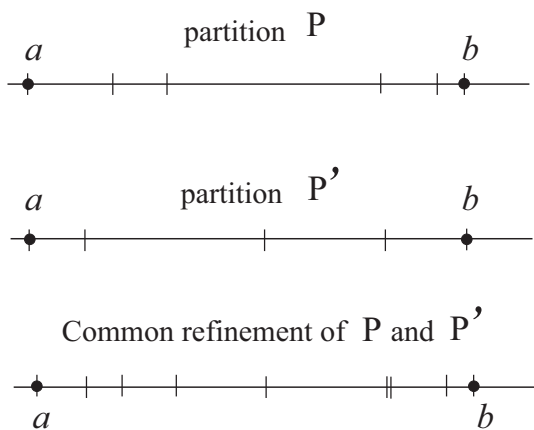
Figure 7.3: The common refinement.

Proof: If f is Riemann integrable then the assertion of the lemma follows immediately from the definition of the integral.

For the converse note that (7.9.1) certainly implies that, if $\epsilon > 0$, then there is a $\delta > 0$ such that, if $\mathcal{P}$ and $\mathcal{P}'$ are partitions of $[a, b]$ with $m(\mathcal{P}) < \delta$ and $m(\mathcal{P}') < \delta$, then

$$|\mathcal{R}(f, \mathcal{P}) - \mathcal{R}(f, \mathcal{P}')| < \epsilon \tag{7.9.2}$$

(just use the triangle inequality).

Now, for each $\epsilon_j = 2^{-j}$, $j = 1, 2, \ldots$, we can choose a $\delta_j > 0$ as in (7.9.2). Let S_j be the *closure* of the set

$$\{\mathcal{R}(f, \mathcal{P}) : m(\mathcal{P}) < \delta_j\}.$$

By the choice of δ_j, the set S_j is contained in a closed interval of length not greater than $2\epsilon_j$.

On the one hand,

$$\bigcap_j S_j$$

must be nonempty since it is the decreasing intersection of compact sets. On the other hand, the length estimate implies that the intersection must be contained in a closed interval of length 0—that is, the intersection is a point. That point is then the limit of the Riemann sums, that is, it is the value of the Riemann integral. $\square$

The most important, and perhaps the simplest, fact about the Riemann integral is that a large class of familiar functions is Riemann integrable.

Theorem 7.10 *Let f be a continuous function on a nontrivial closed, bounded interval $I = [a, b]$. Then f is Riemann integrable on $[a, b]$.*

Proof: We use the lemma. Given $\epsilon > 0$, choose (by the uniform continuity of f on I—Theorem 5.27) a $\delta > 0$ such that, whenever $|s - t| < \delta$ then

$$|f(s) - f(t)| < \frac{\epsilon}{b - a}. \tag{7.10.1}$$

Let $\mathcal{P}$ and $\mathcal{P}'$ be any two partitions of $[a, b]$ of mesh smaller than δ. Let $\mathcal{Q}$ be the common refinement of $\mathcal{P}$ and $\mathcal{P}'$.

Now we let I_j denote the intervals arising in the partition $\mathcal{P}$ (and having length Δ_j) and $\widetilde{I}_\ell$ the intervals arising in the partition $\mathcal{Q}$ (and having length $\widetilde{\Delta}_\ell$). Since the partition $\mathcal{Q}$ contains every point of $\mathcal{P}$, plus some additional points as well, every $\widetilde{I}_\ell$ is contained in some I_j. Fix j and consider the expression

$$\left| f(s_j)\Delta_j - \sum_{\widetilde{I}_\ell \subseteq I_j} f(t_\ell)\widetilde{\Delta}_\ell \right|. \tag{7.10.2}$$

We write

$$\Delta_j = \sum_{\widetilde{I}_\ell \subseteq I_j} \widetilde{\Delta}_\ell.$$

This equality enables us to rearrange (7.10.2) as

$$\left| f(s_j) \cdot \sum_{\widetilde{I}_\ell \subseteq I_j} \widetilde{\Delta}_\ell - \sum_{\widetilde{I}_\ell \subseteq I_j} f(t_\ell)\widetilde{\Delta}_\ell \right|$$

$$= \left| \sum_{\widetilde{I}_\ell \subseteq I_j} [f(s_j) - f(t_\ell)]\widetilde{\Delta}_\ell \right|$$

$$\leq \sum_{\widetilde{I}_\ell \subseteq I_j} |f(s_j) - f(t_\ell)|\widetilde{\Delta}_\ell.$$

But each of the points t_ℓ is in the interval I_j, as is s_j. So they differ by less than δ. Therefore, by (7.10.1), the last expression is less than

$$\sum_{\widetilde{I}_\ell \subseteq I_j} \frac{\epsilon}{b - a}\widetilde{\Delta}_\ell \;=\; \frac{\epsilon}{b - a} \sum_{\widetilde{I}_\ell \subseteq I_j} \widetilde{\Delta}_\ell$$

$$= \frac{\epsilon}{b - a} \cdot \Delta_j.$$

Now we conclude the argument by writing

$$
\begin{aligned}
|\mathcal{R}(f,\mathcal{P}) - \mathcal{R}(f,\mathcal{Q})| \;&=\; \left| \sum_j f(s_j)\Delta_j - \sum_\ell f(t_\ell)\tilde{\Delta}_\ell \right| \\
&\leq\; \sum_j \left| f(s_j)\Delta_j - \sum_{\tilde{I}_\ell \subseteq I_j} f(t_\ell)\tilde{\Delta}_\ell \right| \\
&<\; \sum_j \frac{\epsilon}{b-a} \cdot \Delta_j \\
&=\; \frac{\epsilon}{b-a} \cdot \sum_j \Delta_j \\
&=\; \frac{\epsilon}{b-a} \cdot (b-a) \\
&=\; \epsilon.
\end{aligned}
$$

The estimate for $|\mathcal{R}(f,\mathcal{P}') - \mathcal{R}(f,\mathcal{Q})|$ is identical and we omit it. The result now follows from Lemma 7.9.   □

In the exercises we will ask you to extend the theorem to the case of functions f on $[a,b]$ that are bounded and have finitely many, or even countably many, discontinuities.

We conclude this section by noting an important fact about Riemann integrable functions. A Riemann integrable function on an interval $[a,b]$ *must be bounded*. If it were not, then one could choose the points s_j in the construction of $\mathcal{R}(f,\mathcal{P})$ so that $f(s_j)$ is arbitrarily large, and the Riemann sums would become arbitrarily large, hence cannot converge. You will be asked in the exercises to work out the details of this assertion.

Exercises

1. If f is a Riemann integrable function on $[a,b]$ then show that f must be a bounded function.

2. Define the *Dirichlet function* to be

$$
f(x) = \begin{cases} 1 & \text{if} \quad x \text{ is rational} \\ 0 & \text{if} \quad x \text{ is irrational} \end{cases}
$$

   Prove that the Dirichlet function is not Riemann integrable on the interval $[a,b]$.

3. Define

$$
g(x) = \begin{cases} x \cdot \sin(1/x) & \text{if} \quad x \neq 0 \\ 0 & \text{if} \quad x = 0 \end{cases}
$$

   Is g Riemann integrable on the interval $[-1,1]$?

4. To what extent is the following statement true? If f is Riemann integrable on $[a, b]$ then $1/f$ is Riemann integrable on $[a, b]$.

5. Show that any Riemann integrable function is the pointwise limit of continuous functions.

6. Write the Riemann sum for the function $f(x) = \sin(\pi x^2)$ on the interval $[1, 3]$ with a partition of five equally spaced points.

7. Prove that, if f is continuous on the interval $[a, b]$ except for finitely many discontinuities of the first kind, and if f is bounded, then f is Riemann integrable on $[a, b]$.

8. Do Exercise **9** with the phrase "finitely many" replaced by "countably many."

9. Provide the details of the assertion that, if f is Riemann integrable on the interval $[a, b]$ then, for any $\epsilon > 0$, there is a $\delta > 0$ such that, if $\mathcal{P}$ is a partition of mesh less than δ, then

$$\sum_j \left(\sup_{I_j} f - \inf_{I_j} f \right) \Delta_j < \epsilon \, .$$

[**Hint:** Follow the scheme presented in Remark 7.7. Given $\epsilon > 0$, choose $\delta > 0$ as in the definition of the integral. Fix a partition $\mathcal{P}$ with mesh smaller than δ. Let $K + 1$ be the number of points in $\mathcal{P}$. Choose points $t_j \in I_j$ so that $|f(t_j) - \sup_{I_j} f| < \epsilon/(2(K+1))$; also choose points $t'_j \in I_j$ so that $|f(t'_j) - \inf_{I_j} f| < \epsilon/(2(K+1))$. By applying the definition of the integral to this choice of t_j and t'_j we find that

$$\sum_j \left(\sup_{I_j} f - \inf_{I_j} f \right) \Delta_j < 2\epsilon \, .$$

The result follows.]

* 10. Give an example of a function f such that f^2 is Riemann integrable but f is not.

7.2 Properties of the Riemann Integral

We begin this section with a few elementary properties of the integral that reflect its linear nature.

Theorem 7.11 *Let $[a, b]$ be a nonempty, bounded interval, let f and g be Riemann integrable functions on the interval, and let α be a real number. Then $f \pm g$ and $\alpha \cdot f$ are integrable and we have*

(a) $\int_a^b f(x) \pm g(x)\, dx = \int_a^b f(x)\, dx \pm \int_a^b g(x)\, dx;$

(b) $\int_a^b \alpha \cdot f(x)\, dx = \alpha \cdot \int_a^b f(x)\, dx.$

Proof: For **(a)**, let

$$A = \int_a^b f(x)\, dx$$

and

$$B = \int_a^b g(x)\, dx\,.$$

Let $\epsilon > 0$. Choose a $\delta_1 > 0$ such that if $\mathcal{P}$ is a partition of $[a, b]$ with mesh less than δ_1 then

$$|\mathcal{R}(f, \mathcal{P}) - A| < \frac{\epsilon}{2}\,.$$

Similarly choose a $\delta_2 > 0$ such that if $\mathcal{P}$ is a partition of $[a, b]$ with mesh less than δ_2 then

$$|\mathcal{R}(f, \mathcal{P}) - B| < \frac{\epsilon}{2}\,.$$

Let $\delta = \min\{\delta_1, \delta_2\}$. If $\mathcal{P}'$ is any partition of $[a, b]$ with $m(\mathcal{P}') < \delta$ then

$$
\begin{aligned}
|\mathcal{R}(f \pm g, \mathcal{P}') - (A \pm B)| &= |\mathcal{R}(f, \mathcal{P}') \pm \mathcal{R}(g, \mathcal{P}') - (A \pm B)| \\
&\leq |\mathcal{R}(f, \mathcal{P}') - A| + |\mathcal{R}(g, \mathcal{P}') - B| \\
&< \frac{\epsilon}{2} + \frac{\epsilon}{2} \\
&= \epsilon.
\end{aligned}
$$

This means that the integral of $f \pm g$ exists and equals $A \pm B$, as we were required to prove.

The proof of **(b)** follows similar lines but is much easier and we leave it as an exercise for you.   □

Theorem 7.12 *If c is a point of the interval $[a, b]$ and if f is Riemann integrable on both $[a, c]$ and $[c, b]$ then f is integrable on $[a, b]$ and*

$$\int_a^c f(x)\, dx + \int_c^b f(x)\, dx = \int_a^b f(x)\, dx\,.$$

Proof: Let us write

$$A = \int_a^c f(x)\, dx$$

and

$$B = \int_c^b f(x)\, dx\,.$$

Now pick $\epsilon > 0$. There is a $\delta_1 > 0$ such that if $\mathcal{P}$ is a partition of $[a, c]$ with mesh less than δ_1 then

$$|\mathcal{R}(f, \mathcal{P}) - A| < \frac{\epsilon}{3}\,.$$

Similarly, choose $\delta_2 > 0$ such that if $\mathcal{P}'$ is a partition of $[c, b]$ with mesh less than δ_2 then

$$|\mathcal{R}(f, \mathcal{P}') - B| < \frac{\epsilon}{3}.$$

Let M be an upper bound for $|f|$ (recall, from the remark at the end of Section 1, that a Riemann integrable function must be bounded). Set $\delta = \min\{\delta_1, \delta_2, \epsilon/(6M)\}$. Now let $\mathcal{V} = \{v_1, \ldots, v_k\}$ be any partition of $[a, b]$ with mesh less than δ. There is a last point v_n which is in $[a, c]$ and a first point v_{n+1} in $[c, b]$. Observe that $\mathcal{P} = \{v_0, \ldots, v_n, c\}$ is a partition of $[a, c]$ with mesh smaller than δ_1 and $\mathcal{P}' = \{c, v_{n+1}, \ldots, v_k\}$ is a partition of $[c, b]$ with mesh smaller than δ_2. Let us rename the elements of $\mathcal{P}$ as $\{p_0, \ldots, p_{n+1}\}$ and the elements of $\mathcal{P}'$ as $\{p_0', \cdots p_{k-n+1}'\}$. Notice that $p_{n+1} = p_0' = c$. For each j let s_j be a point chosen in the interval $I_j = [v_{j-1}, v_j]$ from the partition $\mathcal{V}$.

Then we have

$$\left| \mathcal{R}(f, \mathcal{V}) - \left[A + B \right] \right|$$

$$= \left| \left(\sum_{j=1}^{n} f(s_j)\Delta_j - A \right) + f(s_{n+1})\Delta_{n+1} + \left(\sum_{j=n+2}^{k} f(s_j)\Delta_j - B \right) \right|$$

$$= \left| \left(\sum_{j=1}^{n} f(s_j)\Delta_j + f(c) \cdot (c - v_n) - A \right) \right.$$

$$+ \left(f(c) \cdot (v_{n+1} - c) + \sum_{j=n+2}^{k} f(s_j)\Delta_j - B \right)$$

$$+ \left. \left(f(s_{n+1}) - f(c) \right) \cdot (c - v_n) + \left(f(s_{n+1}) - f(c) \right) \cdot (v_{n+1} - c) \right|$$

$$\leq \left| \left(\sum_{j=1}^{n} f(s_j)\Delta_j + f(c) \cdot (c - v_n) - A \right) \right|$$

$$+ \left| \left(f(c) \cdot (v_{n+1} - c) + \sum_{j=n+2}^{k} f(s_j)\Delta_j - B \right) \right|$$

$$+ \left| (f(s_{n+1}) - f(c)) \cdot (v_{n+1} - v_n) \right|$$

$$= \left| \mathcal{R}(f, \mathcal{P}) - A \right| + \left| \mathcal{R}(f, \mathcal{P}') - B \right|$$

$$+ \left| (f(s_{n+1}) - f(c)) \cdot (v_{n+1} - v_n) \right|$$

$$< \frac{\epsilon}{3} + \frac{\epsilon}{3} + 2M \cdot \delta$$

$$\leq \epsilon$$

by the choice of δ.

This shows that f is integrable on the entire interval $[a, b]$ and the value of the integral is

$$A + B = \int_a^c f(x)\, dx + \int_c^b f(x)\, dx. \qquad \square$$

Remark 7.13 The last proof illustrates why it is useful to be able to choose the $s_j \in I_j$ arbitrarily.

EXAMPLE 7.14 If we adopt the convention that

$$\int_b^a f(x)\, dx = -\int_a^b f(x)\, dx$$

(which is consistent with the way that the integral was defined in the first place), then Theorem 7.12 is true even when c is not an element of $[a, b]$. For instance, suppose that $c < a < b$. Then, by Theorem 7.12,

$$\int_c^a f(x)\, dx + \int_a^b f(x)\, dx = \int_c^b f(x)\, dx.$$

But this may be rearranged to read

$$\int_a^b f(x)\, dx = -\int_c^a f(x)\, dx + \int_c^b f(x)\, dx = \int_a^c f(x)\, dx + \int_c^b f(x)\, dx. \qquad \square$$

One of the basic tools of analysis is to perform estimates. Thus we require certain fundamental inequalities about integrals. These are recorded in the next theorem.

Theorem 7.15 Let f and g be integrable functions on a nonempty interval $[a, b]$. Then

(i) $\left| \int_a^b f(x)\, dx \right| \leq \int_a^b |f(x)|\, dx;$

(ii) *If $f(x) \leq g(x)$ for all $x \in [a, b]$ then $\int_a^b f(x)\, dx \leq \int_a^b g(x)\, dx$.*

Proof: If $\mathcal{P}$ is any partition of $[a, b]$ then

$$|\mathcal{R}(f, \mathcal{P})| \leq \mathcal{R}(|f|, \mathcal{P})\,.$$

The first assertion follows.

Next, for part **(ii)**,
$$\mathcal{R}(f, \mathcal{P}) \leq \mathcal{R}(g, \mathcal{P})\,.$$

This inequality implies the second assertion.   □

EXAMPLE 7.16 We may estimate the integral

$$\int_0^1 \sin^3 x\, dx$$

as follows. We apply Theorem 7.15(i) to see that

$$\left| \int_0^1 \sin^3 x\, dx \right| \leq \int_0^1 |\sin^3 x|\, dx\,.$$

Now we apply Theorem 7.15(ii) to determine that

$$\int_0^1 |\sin^3 x|\, dx \leq \int_0^1 1\, dx = 1\,.$$ □

Exercises

1. Suppose that f is a continuous, nonnegative function on the interval $[0, 1]$. Let M be the maximum of f on the interval. Prove that

$$\lim_{n \to \infty} \left[\int_0^1 f(t)^n\, dt \right]^{1/n} = M\,.$$

2. Let f be a bounded function on an unbounded interval of the form $[A, \infty)$. We say that f is integrable on $[A, \infty)$ if f is integrable on every compact subinterval of $[A, \infty)$ and

$$\lim_{B \to +\infty} \int_A^B f(x)\, dx$$

exists and is finite.

Assume that f is nonnegative and Riemann integrable on $[1, N]$ for every $N > 1$ and that f is decreasing. Show that f is Riemann integrable on $[1, \infty)$ if and only if $\sum_{j=1}^\infty f(j)$ is finite.

Suppose that g is nonnegative and integrable on $[1, \infty)$. If $0 \leq |f(x)| \leq g(x)$ for $x \in [1, \infty)$, and f is integrable on compact subintervals of $[1, \infty)$, then prove that f is integrable on $[1, \infty)$.

3. Let f be a function on an interval of the form $(a, b]$ such that f is integrable on compact subintervals of $(a, b]$. If

$$\lim_{\epsilon \to 0^+} \int_{a+\epsilon}^b f(x)\, dx$$

exists and is finite then we say that f is integrable on $(a, b]$. Prove that, if we restrict attention to bounded f, then in fact this definition gives rise to no new integrable functions. However, there are unbounded functions that can now be integrated. Give an example.

Give an example of a function g that is integrable by the definition in the preceding paragraph but is such that $|g|$ is not integrable.

4. If $\int_3^6 f(x)\, dx = 2$ and $\int_3^8 f(x)\, dx = 5$, then calculate $\int_8^6 f(x)\, dx$.

5. Fix a continuous function g on the interval $[0, 1]$. Define

$$Tf = \int_0^1 f(x)g(x)\, dx$$

for f integrable on $[0, 1]$. Prove that

$$|Tf| \le C \int_0^1 |f(x)|\, dx\,.$$

What does the constant C depend on?

6. Let f and g be continuous functions on the interval $[a, b]$. Prove that

$$\int_a^b |f(x) \cdot g(x)|\, dx \le \int_a^b |f(x)|^2\, dx^{1/2} \cdot \int_a^b |g(x)|^2\, dx^{1/2}\,.$$

7. Prove part (b) of Theorem 7.11.

$*$ **8.** Let $1 < p < \infty$ and $q = p/(p-1)$. Let f and g be continuous functions on the interval $[a, b]$. Prove that

$$\int_a^b |f(x) \cdot g(x)|\, dx \le \int_a^b |f(x)|^p\, dx^{1/p} \cdot \int_a^b |g(x)|^q\, dx^{1/q}\,.$$

$*$ **9.** Prove that

$$\lim_{\eta \to 0^+} \int_\eta^{1/\eta} \frac{\cos(2r) - \cos r}{r}\, dr$$

exists.

$*$ **10.** Suppose that f is a Riemann integrable function on the interval $[0, 1]$. Let $\epsilon > 0$. Show that there is a polynomial p so that

$$\int_0^1 |f(x) - p(x)|\, dx < \epsilon\,.$$

1

7.3 Change of Variable and Related Ideas

Another fundamental operation in the theory of the integral is "change of variable" (sometimes called the "u-substitution" in calculus books). We next turn to a careful formulation and proof of this operation. First we need a lemma:

Lemma 7.17 *If f is a Riemann integrable function on $[a, b]$ and if ϕ is a continuous function on a compact interval that contains the range of f then $\phi \circ f$ is Riemann integrable.*

Proof: Let $\epsilon > 0$. Since ϕ is a continuous function on a compact set, it is uniformly continuous (Theorem 5.27). Let $\delta > 0$ be selected such that **(i)** $\delta < \epsilon$ and **(ii)** if $|x - y| < \delta$ then $|\phi(x) - \phi(y)| < \epsilon$.

Now the hypothesis that f is Riemann integrable implies that there exists a $\widetilde{\delta} > 0$ such that if $\mathcal{P}$ and $\mathcal{P}'$ are partitions of $[a, b]$ and $m(\mathcal{P}), m(\mathcal{P}') < \widetilde{\delta}$ then (by Lemma 7.9), for the common refinement $\mathcal{Q}$ of $\mathcal{P}$ and $\mathcal{P}'$, it holds that

$$|\mathcal{R}(f, \mathcal{P}) - \mathcal{R}(f, \mathcal{Q})| < \delta^2 \quad \text{and} \quad \left|\mathcal{R}(f, \widetilde{\mathcal{P}}) - \mathcal{R}(f, \mathcal{Q})\right| < \delta^2 .$$

Fix such a $\mathcal{P}, \mathcal{P}'$ and $\mathcal{Q}$. Let J_ℓ be the intervals of $\mathcal{Q}$ and I_j the intervals of $\mathcal{P}$. Each J_ℓ is contained in some $I_{j(\ell)}$. We write

$$
\begin{aligned}
&\left|\mathcal{R}(\phi \circ f, \mathcal{P}) - \mathcal{R}(\phi \circ f, \mathcal{Q})\right| \\
&= \left|\sum_j \phi \circ f(t_j)\Delta_j - \sum_\ell \phi \circ f(s_\ell)\Delta_\ell\right| \\
&= \left|\sum_j \sum_{J_\ell \subseteq I_j} \phi \circ f(t_j)\Delta_\ell - \sum_j \sum_{J_\ell \subseteq I_j} \phi \circ f(s_\ell)\Delta_\ell\right| \\
&= \left|\sum_j \sum_{J_\ell \subseteq I_j} \left[\phi \circ f(t_j) - \phi \circ f(s_\ell)\right]\Delta_\ell\right| \\
&\leq \left|\sum_j \sum_{J_\ell \subseteq I_j, \ell \in G} \left[\phi \circ f(t_j) - \phi \circ f(s_\ell)\right]\Delta_\ell\right| \\
&\quad + \left|\sum_j \sum_{J_\ell \subseteq I_j, \ell \in B} \left[\phi \circ f(t_j) - \phi \circ f(s_\ell)\right]\Delta_\ell\right|,
\end{aligned}
$$

where we put ℓ in G if $J_\ell \subseteq I_{j(\ell)}$ and $0 \leq \left(\sup_{I_{j(\ell)}} f - \inf_{I_{j(\ell)}} f \right) < \delta$; otherwise we put ℓ into B. Notice that

$$
\begin{aligned}
\sum_{\ell \in B} \delta \Delta_\ell &\leq \sum_{\ell \in B} \left(\sup_{I_{j(\ell)}} f - \inf_{I_{j(\ell)}} f \right) \cdot \Delta_\ell \\
&= \sum_{j=1}^{k} \sum_{J_\ell \subseteq I_j} \left(\sup_{I_j} f - \inf_{I_j} f \right) \cdot \Delta_\ell \\
&= \sum_{j=1}^{k} \left(\sup_{I_j} f - \inf_{I_j} f \right) \Delta_j \\
&< \delta^2
\end{aligned}
$$

by the choice of $\widetilde{\delta}$ (and Remark 7.7). Therefore

$$
\sum_{\ell \in B} \Delta_\ell < \delta \,.
$$

Let M be an upper bound for $|\phi|$ (Corollary 5.22). Then

$$
\begin{aligned}
\left| \sum_{j} \sum_{J_\ell \subseteq I_j, \ell \in B} \left(\phi \circ f(t_j) - \phi \circ f(s_\ell) \right) \Delta_\ell \right| &\leq \left| \sum_{j} \sum_{J_\ell \subseteq I_j, \ell \in B} \left(2 \cdot M \right) \Delta_\ell \right| \\
&\leq 2 \cdot \delta \cdot M \\
&< 2M\epsilon.
\end{aligned}
$$

Also

$$
\left| \sum_{j} \sum_{J_\ell \subseteq I_j, \ell \in G} \left(\phi \circ f(t_j) - \phi \circ f(s_\ell) \right) \Delta_\ell \right| \leq \left| \sum_{j} \sum_{J_\ell \subseteq I_j, \ell \in G} \epsilon \Delta_\ell \right|
$$

since, for $\ell \in G$, we know that $|f(\alpha) - f(\beta)| < \delta$ for any $\alpha, \beta \in I_{j(\ell)}$. However, the last line does not exceed $(b - a) \cdot \epsilon$. Putting together our estimates, we find that

$$
|\mathcal{R}(\phi \circ f, \mathcal{P}) - \mathcal{R}(\phi \circ f, \mathcal{Q})| < \epsilon \cdot (2M + (b - a)) \,.
$$

By symmetry, an analogous inequality holds for $\mathcal{P}'$. By Lemma 7.9, this is what we needed to prove. $\square$

An easier result is that, if f is Riemann integrable on an interval $[a, b]$ and if $\mu : [\alpha, \beta] \to [a, b]$ is continuously differentiable, then $f \circ \mu$ is Riemann integrable (see the exercises).

Corollary 7.18 *If f and g are Riemann integrable on $[a, b]$, then so is the function $f \cdot g$.*

Proof: By Theorem 7.11, $f + g$ is integrable. By the lemma, $(f + g)^2 = f^2 + 2f \cdot g + g^2$ is integrable. But the lemma also implies that f^2 and g^2 are integrable (here we use the function $\phi(x) = x^2$). It results, by subtraction, that $2 \cdot f \cdot g$ is integrable. Hence $f \cdot g$ is integrable. $\qquad\square$

Theorem 7.19 *Let f be an integrable function on an interval $[a, b]$ of positive length. Let ψ be a continuously differentiable function from another interval $[\alpha, \beta]$ of positive length into $[a, b]$. Assume that ψ is increasing, one-to-one, and onto. Then*

$$\int_a^b f(x)\, dx = \int_\alpha^\beta f(\psi(x)) \cdot \psi'(x)\, dx\,.$$

Proof: Since f is integrable, its absolute value is bounded by some number M. Fix $\epsilon > 0$. Since ψ' is continuous on the compact interval $[\alpha, \beta]$, it is uniformly continuous (Theorem 5.27). Hence we may choose $\delta > 0$ so small that if $|s - t| < \delta$ then $|\psi'(s) - \psi'(t)| < \epsilon/(M \cdot (\beta - \alpha))$. If $\mathcal{P} = \{p_0, \ldots, p_k\}$ is any partition of $[a, b]$ then there is an associated partition $\widetilde{\mathcal{P}} = \{\psi^{-1}(p_0), \ldots, \psi^{-1}(p_k)\}$ of $[\alpha, \beta]$. For simplicity denote the points of $\widetilde{\mathcal{P}}$ by $\widetilde{p}_j$. Let us choose the partition $\mathcal{P}$ so fine that the mesh of $\widetilde{\mathcal{P}}$ is less than δ. If t_j are points of $I_j = [p_{j-1}, p_j]$ then there are corresponding points $s_j = \psi^{-1}(t_j)$ of $\widetilde{I}_j = [\widetilde{p}_{j-1}, \widetilde{p}_j]$. Then we have

$$\begin{aligned}
\sum_{j=1}^k f(t_j)\Delta_j &= \sum_{j=1}^k f(t_j)(p_j - p_{j-1}) \\
&= \sum_{j=1}^k f(\psi(s_j))(\psi(\widetilde{p}_j) - \psi(\widetilde{p}_{j-1})) \\
&= \sum_{j=1}^k f(\psi(s_j))\psi'(u_j)(\widetilde{p}_j - \widetilde{p}_{j-1}),
\end{aligned}$$

where we have used the Mean Value Theorem in the last line to find each u_j. Our problem at this point is that $f \circ \psi$ and ψ' are evaluated at different points. So we must do some estimation to correct that problem.

The last displayed line equals

$$\sum_{j=1}^k f(\psi(s_j))\psi'(s_j)(\widetilde{p}_j - \widetilde{p}_{j-1}) + \sum_{j=1}^k f(\psi(s_j))\left(\psi'(u_j) - \psi'(s_j)\right)(\widetilde{p}_j - \widetilde{p}_{j-1})\,.$$

The first sum is a Riemann sum for $f(\psi(x)) \cdot \psi'(x)$ and the second sum is an error term. Since the points u_j and s_j are elements of the same interval $\widetilde{I}_j$ of length less than δ, we conclude that $|\psi'(u_j) - \psi'(s_j)| < \epsilon/(M \cdot |\beta - \alpha|)$. Thus the error term in absolute value does not exceed

$$\sum_{j=1}^k M \cdot \frac{\epsilon}{M \cdot |\beta - \alpha|} \cdot (\widetilde{p}_j - \widetilde{p}_{j-1}) = \frac{\epsilon}{\beta - \alpha} \sum_{j=0}^k (\widetilde{p}_j - \widetilde{p}_{j-1}) = \epsilon\,.$$

This shows that every Riemann sum for f on $[a, b]$ with sufficiently small mesh corresponds to a Riemann sum for $f(\psi(x)) \cdot \psi'(x)$ on $[\alpha, \beta]$ plus an error term of size less than ϵ. A similar argument shows that every Riemann sum for $f(\psi(x)) \cdot \psi'(x)$ on $[\alpha, \beta]$ with sufficiently small mesh corresponds to a Riemann sum for f on $[a, b]$ plus an error term of magnitude less than ϵ. The conclusion is then that the integral of f on $[a, b]$ (which exists by hypothesis) and the integral of $f(\psi(x)) \cdot \psi'(x)$ on $[\alpha, \beta]$ (which exists by the corollary to the lemma) agree. $\square$

EXAMPLE 7.20 Let us analyze the integral

$$\int_0^1 \sin(x^3 + x) \cdot (3x^2 + 1)\, dx \,.$$

We let $f(t) = \sin t$ and $\psi(x) = x^3 + x$. Then we see that the integral has the form

$$\int_0^1 f \circ \psi(x) \cdot \psi'(x)\, dx \,.$$

Here $\psi : [0, 1] \to [0, 2]$.

By the theorem, this integral is equal to

$$\int_0^2 f(t)\, dt = \int_0^2 \sin t\, dt \,. \qquad\qquad \square$$

We conclude this section with the very important

Theorem 7.21 (Fundamental Theorem of Calculus) *Let f be an integrable function on the interval $[a, b]$. For $x \in [a, b]$ we define*

$$F(x) = \int_a^x f(s)\, ds \,.$$

If f is continuous at $x \in (a, b)$ then

$$F'(x) = f(x) \,.$$

Proof: Fix $x \in (a, b)$. Let $\epsilon > 0$. Choose, by the continuity of f at x, a $\delta > 0$ such that $|s - x| < \delta$ implies $|f(s) - f(x)| < \epsilon$. We may assume that $\delta < \min\{x - a, b - x\}$. If $|t - x| < \delta$ then

$$\left| \frac{F(t) - F(x)}{t - x} - f(x) \right| = \left| \frac{\int_a^t f(s)\, ds - \int_a^x f(s)\, ds}{t - x} - f(x) \right|$$

$$= \left| \frac{\int_x^t f(s)\, ds}{t - x} - \frac{\int_x^t f(x)\, ds}{t - x} \right|$$

$$= \left| \frac{\int_x^t (f(s) - f(x))\, ds}{t - x} \right| \,.$$

Notice that we rewrote $f(x)$ as the integral with respect to a dummy variable s over an interval of length $|t - x|$ divided by $(t - x)$. Assume for the moment that $t > x$. Then the last line is dominated by

$$\frac{\int_x^t |f(s) - f(x)|\, ds}{t - x} \leq \frac{\int_x^t \epsilon\, ds}{t - x}$$
$$= \epsilon.$$

A similar estimate holds when $t < x$ (simply reverse the limits of integration).

This shows that

$$\lim_{t \to x} \frac{F(t) - F(x)}{t - x}$$

exists and equals $f(x)$. Thus $F'(x)$ exists and equals $f(x)$. $\qquad\square$

In the exercises we shall consider how to use the theory of one-sided limits to make the conclusion of the Fundamental Theorem true on the entire interval $[a, b]$. We conclude with

Corollary 7.22 *If f is a continuous function on $[a, b]$ and if G is any continuously differentiable function on $[a, b]$ whose derivative equals f on (a, b) then*

$$\int_a^b f(x)\, dx = G(b) - G(a)\,.$$

Proof: Define F as in the theorem. Since F and G have the same derivative on (a, b), they differ by a constant (Corollary 6.18). Then

$$\int_a^b f(x)\, dx = F(b) = F(b) - F(a) = G(b) - G(a)$$

as desired. $\qquad\square$

EXAMPLE 7.23 Let

$$f(x) = \int_0^{x^2} \cos(e^t)\, dt\,.$$

What is the derivative of f?

It is not possible to actually evaluate the given integral, but we can still answer the question. Let $g(s) = \int_0^s \cos(e^t)\, dt$ and let $h(x) = x^2$. Then $f = g \circ h$. Therefore

$$f'(x) = g'(h(x)) \cdot h'(x)\,.$$

Now the Fundamental Theorem of Calculus tells us that

$$g'(s) = \cos(e^s)\,.$$

And obviously $h'(x) = 2x$.

In conclusion,

$$f'(x) = \cos(e^{x^2}) \cdot 2x\,. \qquad\square$$

Exercises

1. Imitate the proof of the Fundamental Theorem of Calculus in this section to show that, if f is continuous on $[a, b]$ and if we define

$$F(x) = \int_a^x f(t)\, dt\,,$$

then the one-sided derivative $F'(a)$ exists and equals $f(a)$ in the sense that

$$\lim_{t \to a^+} \frac{F(t) - F(a)}{t - a} = f(a)\,.$$

Formulate and prove an analogous statement for the one-sided derivative of F at b.

2. Let f be a continuously differentiable function on the interval $[0, 2\pi]$. Further assume that $f(0) = f(2\pi)$ and $f'(0) = f'(2\pi)$. For $n \in \mathbb{N}$ define

$$\widehat{f}(n) = \frac{1}{2\pi} \int_0^{2\pi} f(x) \sin nx\, dx\,.$$

Prove that

$$\sum_{n=1}^\infty |\widehat{f}(n)|^2$$

converges. [**Hint:** Use integration by parts to obtain a favorable estimate on $|\widehat{f}(n)|$.]

3. Let $f_1, f_2, \ldots$ be Riemann integrable functions on $[0, 1]$. Suppose that $f_1(x) \geq f_2(x) \geq \cdots$ for every x and that $\lim_{j \to \infty} f_j(x) \equiv f(x)$ exists and is finite for every x. Is it the case that f is Riemann integrable?

4. Give an example of a function f that is not Riemann integrable but such that f^2 *is* Riemann integrable.

5. Prove that if f is Riemann integrable on the interval $[a, b]$, then f^2 is Riemann integrable on $[a, b]$.

6. Define

$$f(x) = \int_0^{\cos x^2} e^{\sin x}\, dx\,.$$

Calculate $f''(x)$.

7. Give three intuitive reasons why differentiation and integration should be inverse operations.

8. Give an example of an integrable function f and a point x_0 so that

$$F(x) = \int_0^x f(t)\, dt$$

is defined but $F'(x_0) \neq f(x_0)$.

9. Calculate the integral

$$\int_0^1 x^2 \, dx$$

using the original definition of the integral using Riemann sums. Now calculate the integral using the Fundamental Theorem of Calculus. Confirm that both of your answers are the same.

10. It would be foolish to think that

$$\int_a^b f(x) \cdot g(x) \, dx = \int_a^b f(x) \, dx \cdot \int_a^b g(x) \, dx \, .$$

Explain why.

7.4   Another Look at the Integral

For many purposes, such as integration by parts, it is natural to formulate the integral in a more general context than we have considered in the first two sections. Our new formulation is called the *Riemann–Stieltjes integral* and is described below.

Fix an interval $[a, b]$ and a monotonically increasing function α on $[a, b]$. If $\mathcal{P} = \{p_0, p_1, \ldots, p_k\}$ is a partition of $[a, b]$, then let $\Delta\alpha_j = \alpha(p_j) - \alpha(p_{j-1})$. Let f be a bounded function on $[a, b]$ and define the *upper Riemann sum* $\mathcal{U}$ of f with respect to α and the *lower Riemann sum* $\mathcal{L}$ of f with respect to α as follows:

$$\mathcal{U}(f, \mathcal{P}, \alpha) = \sum_{j=1}^k M_j \Delta\alpha_j$$

and

$$\mathcal{L}(f, \mathcal{P}, \alpha) = \sum_{j=1}^k m_j \Delta\alpha_j \, .$$

Here the notation M_j denotes the supremum of f on the interval $I_j = [p_{j-1}, p_j]$ and m_j denotes the infimum of f on I_j.

In the special case $\alpha(x) = x$ the Riemann sums discussed here have a form similar to the Riemann sums considered in the first two sections. Moreover,

$$\mathcal{L}(f, \mathcal{P}, \alpha) \leq \mathcal{R}(f, \mathcal{P}) \leq \mathcal{U}(f, \mathcal{P}, \alpha) \, .$$

We define

$$I^*(f) = \inf \mathcal{U}(f, \mathcal{P}, \alpha)$$

and

$$I_*(f) = \sup \mathcal{L}(f, \mathcal{P}, \alpha) \, .$$

Here the supremum and infimum are taken with respect to all partitions of the interval $[a, b]$. These are, respectively, the *upper* and *lower integrals* of f with respect to α on $[a, b]$.

By definition it is always true that, for any partition $\mathcal{P}$,

$$\mathcal{L}(f, \mathcal{P}, \alpha) \leq I_*(f) \leq I^*(f) \leq \mathcal{U}(f, \mathcal{P}, \alpha).$$

It is natural to declare the integral to exist when the upper and lower integrals agree:

Definition 7.24 Let α be an increasing function on the interval $[a, b]$ and let f be a bounded function on $[a, b]$. We say that the *Riemann–Stieltjes integral of f with respect to α* exists if

$$I^*(f) = I_*(f).$$

When the integral exists we denote it by

$$\int_a^b f \, d\alpha.$$

Notice that the definition of Riemann–Stieltjes integral is different from the definition of Riemann integral that we used in the preceding sections. It turns out that, when $\alpha(x) = x$, the two definitions are equivalent (this assertion is explored in the exercises). In the present generality it is easier to deal with upper and lower integrals in order to determine the existence of integrals.

Definition 7.25 Let $\mathcal{P}$ and $\mathcal{Q}$ be partitions of the interval $[a, b]$. If each point of $\mathcal{P}$ is also an element of $\mathcal{Q}$ then we call $\mathcal{Q}$ a *refinement* of $\mathcal{P}$.

Notice that the refinement $\mathcal{Q}$ is obtained by adding points to $\mathcal{P}$. The mesh of $\mathcal{Q}$ will be less than or equal to that of $\mathcal{P}$. The following lemma enables us to deal effectively with our new language.

Lemma 7.26 Let $\mathcal{P}$ be a partition of the interval $[a, b]$ and f a function on $[a, b]$. Fix an increasing function α on $[a, b]$. If $\mathcal{Q}$ is a refinement of $\mathcal{P}$ then

$$\mathcal{U}(f, \mathcal{Q}, \alpha) \leq \mathcal{U}(f, \mathcal{P}, \alpha)$$

and

$$\mathcal{L}(f, \mathcal{Q}, \alpha) \geq \mathcal{L}(f, \mathcal{P}, \alpha).$$

Proof: Since $\mathcal{Q}$ is a refinement of $\mathcal{P}$ it holds that any interval I_ℓ arising from $\mathcal{Q}$ is contained in some interval $J_{j(\ell)}$ arising from $\mathcal{P}$. Let M_{I_ℓ} be the supremum of f on I_ℓ and $M_{J_{j(\ell)}}$ the supremum of f on the interval $J_{j(\ell)}$. Then $M_{I_\ell} \leq M_{J_{j(\ell)}}$. We conclude that

$$\mathcal{U}(f, \mathcal{Q}, \alpha) = \sum_\ell M_{I_\ell} \Delta \alpha_\ell \leq \sum_\ell M_{J_{j(\ell)}} \Delta \alpha_\ell.$$

We rewrite the right-hand side as

$$\sum_j M_{J_j} \left(\sum_{I_\ell \subseteq J_j} \Delta \alpha_\ell \right).$$

However, because α is monotone, the inner sum simply equals $\alpha(p_j) - \alpha(p_{j-1}) = \Delta\alpha_j$. Thus the last expression is equal to $\mathcal{U}(f, \mathcal{P}, \alpha)$, as desired. In conclusion, $\mathcal{U}(f, \mathcal{Q}, \alpha) \leq \mathcal{U}(f, \mathcal{P}, \alpha)$.

A similar argument applies to the lower sums. $\qquad\blacksquare$

EXAMPLE 7.27 Let $[a, b] = [0, 10]$ and let $\alpha(x)$ be the *greatest integer function*.[1] That is, $\alpha(x)$ is the greatest integer that does not exceed x. So, for example, $\alpha(0.5) = 0, \alpha(2) = 2$, and $\alpha(-3/2) = -2$. Then α is an increasing function on $[0, 10]$. Let f be any continuous function on $[0, 10]$. We shall determine whether

$$\int_0^{10} f \, d\alpha$$

exists and, if it does, calculate its value.

Let $\mathcal{P}$ be a partition of $[0, 10]$. By the lemma, it is to our advantage to assume that the mesh of $\mathcal{P}$ is smaller than 1. Observe that $\Delta\alpha_j$ equals the number of integers that lie in the interval I_j—that is, either 0 or 1. Let $I_{j_0}, I_{j_2}, \ldots I_{j_{10}}$ be, in sequence, the intervals from the partition which do in fact contain each distinct integer (the first of these contains 0, the second contains 1, and so on up to 10). Then

$$\mathcal{U}(f, \mathcal{P}, \alpha) = \sum_{\ell=0}^{10} M_{j_\ell} \Delta\alpha_{j_\ell} = \sum_{\ell=1}^{10} M_{j_\ell}$$

and

$$\mathcal{L}(f, \mathcal{P}, \alpha) = \sum_{\ell=0}^{10} m_{j_\ell} \Delta\alpha_{j_\ell} = \sum_{\ell=1}^{10} m_{j_\ell}$$

because any term in these sums corresponding to an interval not containing an integer must have $\Delta\alpha_j = 0$. Notice that $\Delta\alpha_{j_0} = 0$ since $\alpha(0) = \alpha(p_1) = 0$.

Let $\epsilon > 0$. Since f is uniformly continuous on $[0, 10]$, we may choose a $\delta > 0$ such that $|s - t| < \delta$ implies that $|f(s) - f(t)| < \epsilon/20$. If $m(\mathcal{P}) < \delta$ then it follows that $|f(\ell) - M_{j_\ell}| < \epsilon/20$ and $|f(\ell) - m_{j_\ell}| < \epsilon/20$ for $\ell = 0, 1, \ldots 10$. Therefore

$$\mathcal{U}(f, \mathcal{P}, \alpha) < \sum_{\ell=1}^{10} \left(f(\ell) + \frac{\epsilon}{20} \right)$$

and

$$\mathcal{L}(f, \mathcal{P}, \alpha) > \sum_{\ell=1}^{10} \left(f(\ell) - \frac{\epsilon}{20} \right).$$

Rearranging these inequalities leads to

$$\mathcal{U}(f, \mathcal{P}, \alpha) < \left(\sum_{\ell=1}^{10} f(\ell) \right) + \frac{\epsilon}{2}$$

---

[1]In many texts the greatest integer in x is denoted by $[x]$. We do not use that notation here because it could get confused with our notation for a closed interval.

and

$$\mathcal{L}(f, \mathcal{P}, \alpha) > \left(\sum_{\ell=1}^{10} f(\ell)\right) - \frac{\epsilon}{2}\,.$$

Thus, since $I_*(f)$ and $I^*(f)$ are trapped between $\mathcal{U}$ and $\mathcal{L}$, we conclude that

$$|I_*(f) - I^*(f)| < \epsilon\,.$$

We have seen that, if the partition is fine enough, then the upper and lower integrals of f with respect to α differ by at most ϵ. It follows that $\int_0^{10} f\, d\alpha$ exists. Moreover,

$$\left| I^*(f) - \sum_{\ell=1}^{10} f(\ell) \right| < \epsilon$$

and

$$\left| I_*(f) - \sum_{\ell=1}^{10} f(\ell) \right| < \epsilon\,.$$

We conclude that

$$\int_0^{10} f\, d\alpha = \sum_{\ell=1}^{10} f(\ell)\,. \qquad\qquad \square$$

The example demonstrates that the language of the Riemann–Stieltjes integral allows us to think of the integral as a generalization of the summation process. This is frequently useful, both philosophically and for practical reasons.

The next result, sometimes called Riemann's lemma, is crucial for proving the existence of Riemann–Stieltjes integrals.

Proposition 7.28 (Riemann's Lemma) *Let α be an increasing function on $[a, b]$ and f a bounded function on the interval. The Riemann–Stieltjes integral of f with respect to α exists if and only if, for every $\epsilon > 0$, there is a partition $\mathcal{P}$ such that*

$$|\mathcal{U}(f, \mathcal{P}, \alpha) - \mathcal{L}(f, \mathcal{P}, \alpha)| < \epsilon\,. \qquad\qquad (7.28.1)$$

Proof: First assume that (7.28.1) holds. Fix $\epsilon > 0$. Since $\mathcal{L} \le I_* \le I^* \le \mathcal{U}$, inequality (7.28.1) implies that

$$|I^*(f) - I_*(f)| < \epsilon\,.$$

But this means that $\int_a^b f\, d\alpha$ exists.

Conversely, assume that the integral exists. Fix $\epsilon > 0$. Choose a partition $\mathcal{Q}_1$ such that

$$|\mathcal{U}(f, \mathcal{Q}_1, \alpha) - I^*(f)| < \epsilon/2\,.$$

Likewise choose a partition $\mathcal{Q}_2$ such that

$$|\mathcal{L}(f, \mathcal{Q}_2, \alpha) - I_*(f)| < \epsilon/2\,.$$

Since $I_*(f) = I^*(f)$ it follows that

$$|\mathcal{U}(f, \mathcal{Q}_1, \alpha) - \mathcal{L}(f, \mathcal{Q}_2, \alpha)| < \epsilon. \qquad (7.28.2)$$

Let $\mathcal{P}$ be the common refinement of $\mathcal{Q}_1$ and $\mathcal{Q}_2$. Then we have, again by Lemma 7.26, that

$$\mathcal{L}(f, \mathcal{Q}_2, \alpha) \leq \mathcal{L}(f, \mathcal{P}, \alpha) \leq \int_a^b f \, d\alpha \leq \mathcal{U}(f, \mathcal{P}, \alpha) \leq \mathcal{U}(f, \mathcal{Q}_1, \alpha).$$

But, by (7.28.2), the expressions on the far left and on the far right of these inequalities differ by less than ϵ. Thus $\mathcal{P}$ satifies the condition (7.28.1). □

We note in passing that the basic properties of the Riemann integral noted in Section 2 (Theorems 7.11 and 7.12) hold without change for the Riemann–Stieltjes integral. The proofs are left as exercises for you (use Riemann's lemma!).

Exercises

1. Define $\beta(x)$ by the condition that $\beta(x) = x + k$ when $k \leq x < k + 1$. Calculate

$$\int_2^6 t^2 \, d\beta(t).$$

2. Let $\alpha(x)$ be the greatest integer function as discussed in the text. Define the "fractional part" function by the formula $\gamma(x) = x - \alpha(x)$. Explain why this function has the name "fractional part." Note that γ is not monotone increasing, but it is at least *piecewise* monotone increasing. So the Riemann–Stieltjes integral with respect to γ still makes sense. Calculate

$$\int_0^5 x \, d\gamma(x).$$

3. If p is a polynomial and $\int_a^b p \, d\alpha = 0$ for every choice of α, then what can you conclude about p?

4. Suppose that α and β are monotonic polynomials on the interval $[a, b]$. If $\int f \, d\alpha = \int f \, d\beta$ for every choice of f, then what can you conclude about α and β?

5. Let $\alpha(x)$ be the greatest integer function and $f(x) = x^2$. Calculate $\int_0^3 f \, d\alpha(x)$.

6. Let $f(x) = \alpha(x)$ be the greatest integer function. Calculate $\int_0^4 f \, d\alpha$.

7. State and prove a version of Theorem 7.11 for Riemann–Stieltjes integrals.

8. State and prove a version of Theorem 7.12 for Riemann–Stieltjes integrals.

9. Let $f(x) = x^2$ and $\alpha(x) = x^3$. Calculate

$$\int_0^\pi f \, d\alpha \, .$$

10. State and prove a result to the effect that, when $\alpha(x) = x$, then the Riemann–Stieltjes integral is equivalent with the classical Riemann integral.

11. Any series can be represented as a Riemann–Stieltjes integral. But the converse is not true. Explain.

7.5   Advanced Results on Integration Theory

We now turn to establishing the existence of certain Riemann–Stieltjes integrals.

Theorem 7.29 *Let f be continuous on $[a, b]$ and assume that α is monotonically increasing. Then*

$$\int_a^b f \, d\alpha$$

exists.

Proof: We may assume that α is nonconstant; otherwise there is nothing to prove.

Pick $\epsilon > 0$. By the uniform continuity of f we may choose a $\delta > 0$ such that if $|s - t| < \delta$ then $|f(s) - f(t)| < \epsilon/(\alpha(b) - \alpha(a))$. Let $\mathcal{P}$ be any partition of $[a, b]$ that has mesh smaller than δ. Then

$$
\begin{aligned}
|\mathcal{U}(f, \mathcal{P}, \alpha) - \mathcal{L}(f, \mathcal{P}, \alpha)| &= \left| \sum_j M_j \Delta\alpha_j - \sum_j m_j \Delta\alpha_j \right| \\
&= \sum_j |M_j - m_j| \, \Delta\alpha_j \\
&< \sum_j \frac{\epsilon}{\alpha(b) - \alpha(a)} \Delta\alpha_j \\
&= \frac{\epsilon}{\alpha(b) - \alpha(a)} \cdot \sum_j \Delta\alpha_j \\
&= \epsilon \, .
\end{aligned}
$$

Here, of course, we have used the monotonicity of α to observe that the last sum collapses to $\alpha(b) - \alpha(a)$. By Riemann's lemma, the proof is complete. $\square$

Notice how simple Riemann's lemma is to use. You may find it instructive to compare the proofs of this section with the rather difficult proofs in Section 2. What we are learning is that a good definition (and accompanying lemma(s)) can, in the end, make everything much clearer and simpler. Now we establish a companion result to the first one.

Theorem 7.30 *If α is an increasing and continuous function on the interval $[a, b]$ and if f is monotonic on $[a, b]$ then $\int_a^b f \, d\alpha$ exists.*

Proof: We may assume that $\alpha(b) > \alpha(a)$ and that f is monotone *increasing*. Let $L = \alpha(b) - \alpha(a)$ and $M = f(b) - f(a)$. Pick $\epsilon > 0$. Choose a positive integer k so that

$$\frac{L \cdot M}{k} < \epsilon \,.$$

Let $p_0 = a$ and choose p_1 to be the first point to the right of p_0 such that $\alpha(p_1) - \alpha(p_0) = L/k$ (this is possible, by the Intermediate Value Theorem, since α is continuous). Continuing, choose p_j to be the first point to the right of p_{j-1} such that $\alpha(p_j) - \alpha(p_{j-1}) = L/k$. This process will terminate after k steps and we will have $p_k = b$. Then $\mathcal{P} = \{p_0, p_1, \ldots, p_k\}$ is a partition of $[a, b]$.

Next observe that, for each j, the value M_j of sup f on I_j is $f(p_j)$ since f is increasing. Similarly the value m_j of inf f on I_j is $f(p_{j-1})$. We find therefore that

$$
\begin{aligned}
\mathcal{U}(f, \mathcal{P}, \alpha) - \mathcal{L}(f, \mathcal{P}, \alpha) &= \sum_{j=1}^{k} M_j \Delta \alpha_j - \sum_{j=1}^{k} m_j \Delta \alpha_j \\
&= \sum_{j=1}^{k} \left((M_j - m_j) \frac{L}{k} \right) \\
&= \frac{L}{k} \sum_{j=1}^{k} (f(x_j) - f(x_{j-1})) \\
&= \frac{L \cdot M}{k} \\
&< \epsilon.
\end{aligned}
$$

Therefore inequality (7.28.1) of Riemann's lemma is satisfied and the integral exists. $\qquad\square$

One of the useful features of Riemann–Stieltjes integration is that it puts integration by parts into a very natural setting. We begin with a lemma.

Lemma 7.31 *Let f be continuous on an interval $[a, b]$ and let g be monotone increasing and continuous on that interval. If G is an antiderivative for g then*

$$\int_a^b f(x) g(x) \, dx = \int_a^b f \, dG \,.$$

Proof: Apply the Mean Value Theorem to the Riemann sums for the integral on the right. $\qquad\square$

Theorem 7.32 (Integration by Parts) *Suppose that both f and g are continuous, increasing functions on the interval $[a, b]$. Let F be an antiderivative for f on $[a, b]$ and G an antiderivative for g on $[a, b]$. Then we have*

$$\int_a^b F\, dG = [F(b) \cdot G(b) - F(a) \cdot G(a)] - \int_a^b G\, dF\,.$$

Proof: Notice that, by the preceding lemma, both integrals exist. Set $P(x) = F(x) \cdot G(x)$. Then P has a continuous derivative on the interval $[a, b]$. Thus the Fundamental Theorem applies and we may write

$$P(b) - P(a) = \int_a^b P'(x)\, dx = [F(b) \cdot G(b) - F(a) \cdot G(a)]\,.$$

Now, writing out P' explicitly, using Leibnitz's Rule for the derivative of a product, we obtain

$$\int_a^b F(x) g(x)\, dx = [F(b)G(b) - F(a)G(a)] - \int_a^b G(x) f(x)\, dx\,.$$

But the lemma allows us to rewrite this equation as

$$\int_a^b F\, dG = [F(b)G(b) - F(a)G(a)] - \int_a^b G(x)\, dF\,. \qquad \square$$

Remark 7.33 The integration by parts formula can also be proved by applying *summation* by parts to the Riemann sums for the integral

$$\int_a^b F\, dG\,.$$

This method is explored in the exercises.

We have already observed that the Riemann–Stieltjes integral

$$\int_a^b f\, d\alpha$$

is linear in f; that is,

$$\int_a^b (f + g)\, d\alpha = \int_a^b f\, d\alpha + \int_a^b g\, d\alpha$$

and

$$\int_a^b c \cdot f\, d\alpha = c \cdot \int_a^b f\, d\alpha$$

when both f and g are Riemann–Stieltjes integrable with respect to α and for any constant c. We also would expect, from the very way that the integral is constructed, that it would be linear in the α entry. But we have not even defined

the Riemann–Stieltjes integral for nonincreasing α. And what of a function α that is the difference of two increasing functions? Such a function need not be monotone. Is it possible to identify which functions α can be decomposed as sums or differences of monotonic functions? It turns out that there is a satisfactory answer to these questions, and we should like to discuss these matters briefly.

Definition 7.34 If α is a monotonically *decreasing* function on $[a, b]$ and f is a function on $[a, b]$ then we define

$$\int_a^b f \, d\alpha = -\int_a^b f \, d(-\alpha)$$

when the right side exists.

The definition exploits the simple observation that if α is decreasing then $-\alpha$ is increasing; hence the preceding theory applies to the function $-\alpha$.

Next we have

Definition 7.35 Let α be a function on $[a, b]$ that can be expressed as

$$\alpha(x) = \alpha_1(x) - \alpha_2(x),$$

where both α_1 and α_2 are increasing. Then, for any f on $[a, b]$, we define

$$\int_a^b f \, d\alpha = \int_a^b f \, d\alpha_1 - \int_a^b f \, d\alpha_2,$$

provided that both integrals on the right exist.

Now, by the very way that we have formulated our definitions, $\int_a^b f \, d\alpha$ is linear in both the f entry and the α entry. But the definitions are not satisfactory unless we can identify those α that can actually occur in the last definition. This leads us to a new class of functions.

Definition 7.36 Let f be a function on the interval $[a, b]$. For $x \in [a, b]$ we define

$$Vf(x) = \sup \sum_{j=1}^k |f(p_j) - f(p_{j-1})|,$$

where the supremum is taken over all partitions $\mathcal{P}$ of the interval $[a, x]$.

If $Vf \equiv Vf(b) < \infty$, then the function f is said to be of *bounded variation* on the interval $[a, b]$. In this circumstance the quantity $Vf(b)$ is called the *total variation* of f on $[a, b]$.

A function of bounded variation has the property that its graph does not have unbounded total oscillation.

EXAMPLE 7.37 Define $f(x) = \sin x$, with domain the interval $[0, 2\pi]$. Let us calculate Vf. Let $\mathcal{P}$ be a partition of $[0, 2\pi]$. Since adding points to the partition only makes the sum

$$\sum_{j=1}^{k} |f(p_j) - f(p_{j-1})|$$

larger (by the triangle inequality), we may as well suppose that

$$\mathcal{P} = \{p_0, p_1, p_2, \ldots, p_k\}$$

contains the points $\pi/2, 3\pi/2$. Assume that $p_{\ell_1} = \pi/2$ and $p_{\ell_2} = 3\pi/2$. Then

$$
\begin{aligned}
\sum_{j=1}^{k} |f(p_j) - f(p_{j-1})| &= \sum_{j=1}^{\ell_1} |f(p_j) - f(p_{j-1})| \\
&+ \sum_{j=\ell_1+1}^{\ell_2} |f(p_j) - f(p_{j-1})| \\
&+ \sum_{j=\ell_2+1}^{k} |f(p_j) - f(p_{j-1})| \, .
\end{aligned}
$$

However, f is increasing on the interval $[0, \pi/2] = [0, p_{\ell_1}]$. Therefore the first sum is just

$$\sum_{j=1}^{\ell_1} f(p_j) - f(p_{j-1}) = f(p_{\ell_1}) - f(p_0) = f(\pi/2) - f(0) = 1 \, .$$

Similarly, f is monotone on the intervals $[\pi/2, 3\pi/2] = [p_{\ell_1}, p_{\ell_2}]$ and $[3\pi/2, 2\pi] = [p_{\ell_2}, p_k]$. Thus the second and third sums equal $f(p_{\ell_1}) - f(p_{\ell_2}) = 2$ and $f(p_k) - f(p_{\ell_2}) = 1$ respectively. It follows that

$$Vf = Vf(2\pi) = 1 + 2 + 1 = 4 \, .$$

Of course $Vf(x)$ for any $x \in [0, 2\pi]$ can be computed by similar means (see the exercises). □

EXAMPLE 7.38 In general, if f is a continuously differentiable function on an interval $[a, b]$ then

$$Vf(x) = \int_a^x |f'(t)| \, dt \, .$$

This assertion will be explored in the exercises. □

EXAMPLE 7.39 The function $f(x) = \cos x$ on the interval $[0, 2\pi]$ is of bounded variation. And in fact

$$Vf = 4$$

because the function goes from 1 down to 0 and then from 0 down to -1 and finally from -1 up to 0 and finally from 0 up to 1.

Alternatively, one can obtain the same answer by calculating the integral

$$Vf = \int_0^{2\pi} |f'(x)|\, dx = \int_0^{2\pi} |\sin x|\, dx\,. \qquad \Box$$

Lemma 7.40 *Let f be a function of bounded variation on the interval $[a, b]$. Then the function Vf is increasing on $[a, b]$.*

Proof: Let $s < t$ be elements of $[a, b]$. Let $\mathcal{P} = \{p_0, p_1, \ldots, p_k\}$ be a partition of $[a, s]$. Then $\widetilde{\mathcal{P}} = \{p_0, p_1, \ldots, p_k, t\}$ is a partition of $[a, t]$ and

$$\sum_{j=1}^{k} |f(p_j) - f(p_{j-1})|$$

$$\leq \sum_{j=1}^{k} |f(p_j) - f(p_{j-1})| + |f(t) - f(p_k)|$$
$$\leq Vf(t).$$

Taking the supremum on the left over all partitions $\mathcal{P}$ of $[a, s]$ yields that

$$Vf(s) \leq Vf(t)\,. \qquad \Box$$

Lemma 7.41 *Let f be a function of bounded variation on the interval $[a, b]$. Then the function $Vf - f$ is increasing on the interval $[a, b]$.*

Proof: Let $s < t$ be elements of $[a, b]$. Pick $\epsilon > 0$. By the definition of Vf we may choose a partition $\mathcal{P} = \{p_0, p_1, \ldots, p_k\}$ of the interval $[a, s]$ such that

$$Vf(s) - \epsilon < \sum_{j=1}^{k} |f(p_j) - f(p_{j-1})|\,. \qquad (7.41.1)$$

But then $\widetilde{\mathcal{P}} = \{p_0, p_1, \ldots, p_k, t\}$ is a partition of $[a, t]$ and we have that

$$\sum_{j=1}^{k} |f(p_j) - f(p_{j-1})| + |f(t) - f(s)| \leq Vf(t)\,.$$

Using (7.41.1), we may conclude that

$$Vf(s) - \epsilon + f(t) - f(s) < \sum_{j=1}^{k} |f(p_j) - f(p_{j-1})| + |f(t) - f(s)| \leq Vf(t)\,.$$

We conclude that

$$Vf(s) - f(s) < Vf(t) - f(t) + \epsilon\,.$$

Since the inequality holds for every $\epsilon > 0$, we see that the function $Vf - f$ is increasing.   □

Now we may combine the last two lemmas to obtain our main result:

Proposition 7.42 *If a function f is of bounded variation on $[a, b]$, then f may be written as the difference of two increasing functions. Conversely, the difference of two increasing functions is a function of bounded variation.*

Proof: If f is of bounded variation write $f = Vf - (Vf - f) \equiv f_1 - f_2$. By the lemmas, both f_1 and f_2 are increasing.

For the converse, assume that $f = f_1 - f_2$ with f_1, f_2 increasing. Then it is easy to see that

$$Vf(b) \le |f_1(b) - f_1(a)| + |f_2(b) - f_2(a)| .$$

Thus f is of bounded variation.   □

Now the main point of this discussion is the following theorem:

Theorem 7.43 *If f is a continuous function on $[a, b]$ and if α is of bounded variation on $[a, b]$ then the integral*

$$\int_a^b f \, d\alpha$$

exists and is finite.

If g is of bounded variation on $[a, b]$ and if β is a continuous function of bounded variation on $[a, b]$ then the integral

$$\int_a^b g \, d\beta$$

exists and is finite.

Proof: Write the function(s) of bounded variation as the difference of increasing functions. Then apply Theorems 7.29 and 7.30.   □

Exercises

1. Prove that the integral

$$\int_0^\infty \frac{\sin x}{x} \, dx$$

exists.

2. Prove that, if f is a continuously differentiable function on the interval $[a, b]$, then

$$Vf = \int_a^b |f'(x)| \, dx \, .$$

[**Hint:** You will prove two inequalities. For one, use the Fundamental Theorem. For the other, use the Mean Value Theorem.]

3. Give an example of a continuous function on the interval $[0, 1]$ that is not of bounded variation.

4. Let β be a nonnegative, increasing function on the interval $[a, b]$. Set $m = \beta(a)$ and $M = \beta(b)$. For any number λ lying between m and M, set $S_\lambda = \{x \in [a, b] : \beta(x) > \lambda\}$. Prove that S_λ must be an interval. Let $\ell(\lambda)$ be the length of S_λ. Then prove that

$$\int_a^b \beta(t)^p \, dt \quad = \quad -\int_m^M s^p \, d\ell(s)$$

$$= \quad \int_0^M \ell(s) \cdot p \cdot s^{p-1} \, ds.$$

5. If φ is a convex function on the real line, then prove that, for f integrable on $[0, 1]$,

$$\varphi \left(\int_0^1 f(x) \, dx \right) \le \int_0^1 \varphi(f(x)) \, dx \, .$$

6. Give an example of a continuously differentiable function on an open interval that is not of bounded variation.

7. Prove that a continuously differentiable function on a compact interval is of bounded variation.

8. Let $f(x) = \sin x$ on the interval $[0, 2\pi]$. Calculate $Vf(x)$ for any $x \in [0, 2\pi]$.

9. Let $f(x) = x^2$ and $\alpha(x) = \sin x$. Calculate

$$\int_0^\pi f \, d\alpha \, .$$

10. Calculate the total variation of the function

$$f(x) = \sin(jx)$$

on the interval $(0, \pi)$.

11. Provide a detailed proof of Lemma 7.31.

Chapter 8

Sequences and Series of Functions

8.1 Partial Sums and Pointwise Convergence

A *sequence of functions* is usually written

$$f_1, f_2, \ldots \quad \text{or} \quad \{f_j\}_{j=1}^{\infty} \ .$$

We will generally assume that the functions f_j all have the same domain S.

Definition 8.1 A sequence of functions $\{f_j\}_{j=1}^{\infty}$ with domain $S \subseteq \mathbb{R}$ is said to *converge pointwise* to a limit function f on S if, for each $x \in S$, the sequence of numbers $\{f_j(x)\}$ converges to $f(x)$.

EXAMPLE 8.2 Define $f_j(x) = x^j$ with domain $S = \{x : 0 \le x \le 1\}$. If $0 \le x < 1$ then $f_j(x) \to 0$. However, $f_j(1) \to 1$. Therefore the sequence f_j converges pointwise to the function

$$f(x) = \begin{cases} 0 & \text{if} \quad 0 \le x < 1 \\ 1 & \text{if} \quad x = 1 \end{cases}$$

See Figure 8.1. We see that, even though the f_j are each continuous, the limit function f is not. □

Here are some of the basic questions that we must ask about a sequence of functions f_j that converges to a function f on a domain S:

(1) If the functions f_j are continuous, then is f continuous?

(2) If the functions f_j are integrable on an interval I, then is f integrable on I?

(3) If f is integrable on I, then does the sequence $\int_I f_j(x)\,dx$ converge to $\int_I f(x)\,dx$?
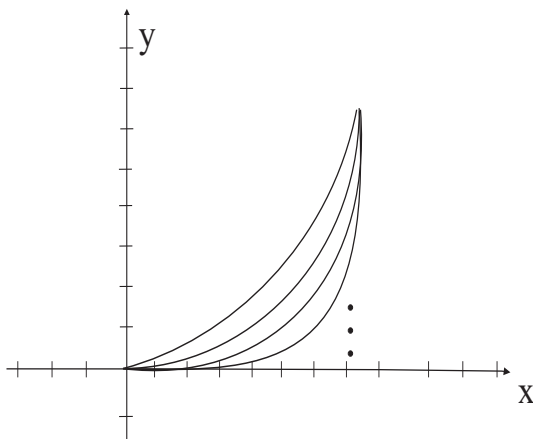
Figure 8.1: The sequence $\{x^j\}$.

(4) If the functions f_j are differentiable, then is f differentiable?

(5) If f is differentiable, then does the sequence f'_j converge to f'?

We see from Example 8.2 that the answer to the first question is "no": Each of the f_j is continuous but f is not. It turns out that, in order to obtain a favorable answer to our questions, we must consider a stricter notion of convergence of functions. This motivates the next definition.

Definition 8.3 Let f_j be a sequence of functions on a domain S. We say that the functions f_j *converge uniformly* to f on S if, given $\epsilon > 0$, there is an $N > 0$ such that, for any $j > N$ and any $x \in S$, it holds that $|f_j(x) - f(x)| < \epsilon$.

Notice that the special feature of uniform convergence is that the rate at which $f_j(x)$ converges is independent of $x \in S$. In Example 8.1, $f_j(x)$ is converging very rapidly to zero for x near zero but arbitrarily slowly to zero for x near 1—see Figure 8.1. In the next example we shall prove this assertion rigorously:

EXAMPLE 8.4 The sequence $f_j(x) = x^j$ does *not* converge uniformly to the limit function
$$f(x) = \begin{cases} 0 & \text{if} \quad 0 \le x < 1 \\ 1 & \text{if} \quad x = 1 \end{cases}$$

on the domain $S = [0, 1]$. In fact it does not even do so on the smaller domain $[0, 1)$. To see this notice that, no matter how large j is, we have by the Mean Value Theorem that

$$f_j(1) - f_j(1 - 1/(2j)) = \frac{1}{2j} \cdot f'_j(\xi)$$

for some ξ between $1 - 1/(2j)$ and 1. But $f_j'(x) = j \cdot x^{j-1}$ hence $|f_j'(\xi)| < j$ and we conclude that

$$|f_j(1) - f_j(1 - 1/(2j))| < \frac{1}{2}$$

or

$$f_j(1 - 1/(2j)) > f_j(1) - \frac{1}{2} = \frac{1}{2}.$$

In conclusion, no matter how large j, there will be values of x (namely, $x = 1 - 1/(2j)$) at which $f_j(x)$ is at least distance $1/2$ from the limit 0. We conclude that the convergence is not uniform. $\qquad\square$

Theorem 8.5 If f_j are continuous functions on a set S that converge uniformly on S to a function f then f is also continuous.

Proof: Let $\epsilon > 0$. Choose an integer N so large that, if $j > N$, then $|f_j(x) - f(x)| < \epsilon/3$ for all $x \in S$. Fix $P \in S$. Choose $\delta > 0$ so small that if $|x - P| < \delta$ then $|f_N(x) - f_N(P)| < \epsilon/3$. For such x we have

$$
\begin{aligned}
|f(x) - f(P)| &\leq |f(x) - f_N(x)| + |f_N(x) - f_N(P)| + |f_N(P) - f(P)| \\
&< \frac{\epsilon}{3} + \frac{\epsilon}{3} + \frac{\epsilon}{3}
\end{aligned}
$$

by the way that we chose N and δ. But the last line sums to ϵ, proving that f is continuous at P. Since $P \in S$ was chosen arbitrarily, we are done. $\qquad\square$

EXAMPLE 8.6 Define functions

$$f_j(x) = \begin{cases} 0 & \text{if} \quad x = 0 \\ j & \text{if} \quad 0 < x \leq 1/j \\ 0 & \text{if} \quad 1/j < x \leq 1 \end{cases}$$

for $j = 2, 3, \ldots$. Then $\lim_{j \to \infty} f_j(x) = 0 \equiv f(x)$ for all x in the interval $I = [0, 1]$. However

$$\int_0^1 f_j(x)\, dx = \int_0^{1/j} j\, dx = 1$$

for every j. Thus the f_j converge to the integrable limit function $f(x) \equiv 0$ (with integral 0), but their integrals do not converge to the integral of f.

Of course the f_j do not converge uniformly. $\qquad\square$

EXAMPLE 8.7 Let $q_1, q_2, \ldots$ be an enumeration of the rationals in the interval $I = [0, 1]$. Define functions

$$f_j(x) = \begin{cases} 1 & \text{if} \quad x \in \{q_1, q_2, \ldots, q_j\} \\ 0 & \text{if} \quad x \notin \{q_1, q_2, \ldots, q_j\} \end{cases}$$

Then the functions f_j converge pointwise to the Dirichlet function f which is equal to 1 on the rationals and 0 on the irrationals. Each of the functions f_j has integral 0 on I. But the function f is not Riemann integrable on I. $\qquad\square$

The last two examples show that something more than pointwise conver-
gence is needed in order for the integral to respect the limit process.

Theorem 8.8 *Let f_j be integrable functions on a nontrivial bounded interval
$[a, b]$ and suppose that the functions f_j converge uniformly to the limit function
f. Then f is integrable on $[a, b]$ and*

$$\lim_{j \to \infty} \int_a^b f_j(x)\, dx = \int_a^b f(x)\, dx\,.$$

Proof: Pick $\epsilon > 0$. Choose N so large that, if $j > N$, then $|f_j(x) - f(x)| <
\epsilon/[2(b - a)]$ for all $x \in [a, b]$. Notice that, if $j, k > N$, then

$$\left| \int_a^b f_j(x)\, dx - \int_a^b f_k(x)\, dx \right| \leq \int_a^b |f_j(x) - f_k(x)|\ dx\,. \qquad (8.8.1)$$

But $|f_j(x) - f_k(x)| \leq |f_j(x) - f(x)| + |f(x) - f_k(x)| < \epsilon/(b - a)$. Therefore line
(8.8.1) does not exceed

$$\int_a^b \frac{\epsilon}{b - a}\, dx = \epsilon\,.$$

Thus the numbers $\int_a^b f_j(x)\, dx$ form a Cauchy sequence. Let the limit of this
sequence be called A. Notice that, if we let $k \to \infty$ in the inequality

$$\left| \int_a^b f_j(x)\, dx - \int_a^b f_k(x)\, dx \right| \leq \epsilon\,,$$

then we obtain

$$\left| \int_a^b f_j(x)\, dx - A \right| \leq \epsilon$$

for all $j \geq N$. This estimate will be used below.

By hypothesis there is a $\delta > 0$ such that, if $\mathcal{P} = \{p_1, \ldots, p_k\}$ is a partition
of $[a, b]$ with $m(\mathcal{P}) < \delta$, then

$$\left| \mathcal{R}(f_N, \mathcal{P}) - \int_a^b f_N(x)\, dx \right| < \epsilon\,.$$

But then, for such a partition, we have

$$|\mathcal{R}(f, \mathcal{P}) - A| \ \leq\ \left| \mathcal{R}(f, \mathcal{P}) - \mathcal{R}(f_N, \mathcal{P}) \right| + \left| \mathcal{R}(f_N, \mathcal{P}) - \int_a^b f_N(x)\, dx \right|$$

$$+ \left| \int_a^b f_N(x)\, dx - A \right|\,.$$

We have already noted that, by the choice of N, the third term on the right is
smaller than ϵ. The second term is smaller than ϵ by the way that we chose the

partition $\mathcal{P}$. It remains to examine the first term. Now

$$
\begin{aligned}
\left| \mathcal{R}(f, \mathcal{P}) - \mathcal{R}(f_N, \mathcal{P}) \right| &= \left| \sum_{j=1}^{k} f(s_j) \Delta_j - \sum_{j=1}^{k} f_N(s_j) \Delta_j \right| \\
&\leq \sum_{j=1}^{k} \left| f(s_j) - f_N(s_j) \right| \Delta_j \\
&< \sum_{j=1}^{k} \frac{\epsilon}{2(b-a)} \Delta_j \\
&= \frac{\epsilon}{2(b-a)} \sum_{j=1}^{k} \Delta_j \\
&= \frac{\epsilon}{2}.
\end{aligned}
$$

Therefore $|\mathcal{R}(f, \mathcal{P}) - A| < 3\epsilon$ when $m(\mathcal{P}) < \delta$. This shows that the function f is integrable on $[a, b]$ and has integral with value A. □

We have succeeded in answering questions **(1)** and **(2)** that were raised at the beginning of the section. In the next section we will answer questions **(3)**, **(4)**, and **(5)**.

EXAMPLE 8.9 Define

$$
f_j(x) = \begin{cases}
0 & \text{if} & x \leq j \\
x - j & \text{if} & j < x \leq j+1 \\
(j+2) - x & \text{if} & j+1 < x \leq j+2 \\
0 & \text{if} & j+2 < x.
\end{cases}
$$

Then

$$
\int f_j(x)\, dx = 1
$$

for each j. But

$$
\lim_{j \to \infty} f_j(x) = 0
$$

for every x. So we see that

$$
1 = \lim_{j \to \infty} \int f_j(x)\, dx \neq \int \lim_{j \to \infty} f_j(x)\, dx = \int 0\, dx = 0. \qquad \square
$$

Exercises

1. If $f_j \to f$ uniformly on a domain S and if f_j, f never vanish on S then does it follow that the functions $1/f_j$ converge uniformly to $1/f$ on S?

2. Write out the first five partial sums for the series

$$\sum_{j=1}^{\infty} \frac{\sin^3 j}{j^2}.$$

3. Write a series of polynomials that converges to $f(x) = \sin x^2$. Can you prove that it converges?

4. Write a series of trigonometric functions that converges to $f(x) = x$. Can you prove that it converges?

5. Write a series of piecewise linear functions that converges to $f(x) = x^2$ on the interval $[0, 1]$. Can you prove that it converges?

6. Write a series of functions that converges pointwise on $[0, 1]$ but does not converge uniformly on any proper subinterval. [**Hint:** First consider a sequence.]

7. Give an example of a Taylor series that converges uniformly on compact sets to its limit function.

8. Prove that the series

$$\sum_{j=1}^{\infty} \frac{\sin jx}{j^2}$$

converges uniformly to a continuous function on the interval $[0, 1]$.

9. A Taylor series will never converge only pointwise. Explain.

10. Define
$$f_j(x) = \begin{cases} 1 + 1/j & \text{if} \quad x < j \\ 1/j & \text{if} \quad x \geq j \end{cases}$$
Show that f_j converges to the identically 1 function pointwise but not uniformly.

11. Define
$$f_j(x) = \begin{cases} 0 & \text{if} \quad x \leq 0 \\ x^2/j & \text{if} \quad x > 0 \end{cases}$$
Prove that each f_j is continuous, and the sequence $\{f_j\}$ converges pointwise to the identically 0 function. But the sequence does not converge uniformly.

12. Show that, if $\sum_j f_j'$ converges uniformly on $[0, 1]$ (where the prime stands for the derivative), and if $f_j(0) = 0$ for all j, then $\sum_j f_j$ converges uniformly on compact sets.

13. TRUE or FALSE: If $\sum_j f_j$ converges absolutely and uniformly and $\sum_j g_j$ converges absolutely and uniformly on a compact interval $[a, b]$, then so does $\sum_j f_j g_j$.

8.2 More on Uniform Convergence

In general, limits do not commute. Since the integral is defined with a limit, and since we saw in the last section that integrals do not always respect limits of functions, we know some concrete instances of non-commutation of limits. The fact that continuity is defined with a limit, and that the limit of continuous functions need not be continuous, gives even more examples of situations in which limits do not commute. Let us now turn to a situation in which limits *do* commute:

Theorem 8.10 *Fix a set S and a point $s \in S$. Assume that the functions f_j converge uniformly on the domain $S \setminus \{s\}$ to a limit function f. Suppose that each function $f_j(x)$ has a limit as $x \to s$. Then f itself has a limit as $x \to s$ and*

$$\lim_{x \to s} f(x) = \lim_{j \to \infty} \lim_{x \to s} f_j(x).$$

Because of the way that f is defined, we may rewrite this conclusion as

$$\lim_{x \to s} \lim_{j \to \infty} f_j(x) = \lim_{j \to \infty} \lim_{x \to s} f_j(x).$$

In other words, the limits $\lim_{x \to s}$ and $\lim_{j \to \infty}$ commute.

Proof: Let $\alpha_j = \lim_{x \to s} f_j(x)$. Let $\epsilon > 0$. There is a number $N > 0$ (independent of $x \in S \setminus \{s\}$) such that $j > N$ implies that $|f_j(x) - f(x)| < \epsilon/4$. Fix $j, k > N$. Choose $\delta > 0$ such that $0 < |x - s| < \delta$ implies both that $|f_j(x) - \alpha_j| < \epsilon/4$ and $|f_k(x) - \alpha_k| < \epsilon/4$. Then

$$|\alpha_j - \alpha_k| \leq |\alpha_j - f_j(x)| + |f_j(x) - f(x)| + |f(x) - f_k(x)| + |f_k(x) - \alpha_k|.$$

The first and last expressions are less than $\epsilon/4$ by the choice of x. The middle two expressions are less than $\epsilon/4$ by the choice of N (and therefore of j and k). We conclude that the sequence α_j is Cauchy. Let α be the limit of that sequence.

Letting $k \to \infty$ in the inequality

$$|\alpha_j - \alpha_k| < \epsilon$$

that we obtained above yields

$$|\alpha_j - \alpha| \leq \epsilon$$

for $j > N$. Now, with δ as above and $0 < |x - s| < \delta$, we have

$$|f(x) - \alpha| \leq |f(x) - f_j(x)| + |f_j(x) - \alpha_j| + |\alpha_j - \alpha|.$$

By the choices we have made, the first term is less than $\epsilon/4$, the second is less than $\epsilon/4$, and the third is less than or equal to ϵ. Altogether, if $0 < |x - s| < \delta$ then $|f(x) - \alpha| < 2\epsilon$. This is the desired conclusion. $\square$

EXAMPLE 8.11 Consider the example

$$f_j(x) = x^j$$

on the interval $[0, 1]$. We see that

$$\lim_{j \to \infty} f_j(x) = 0 \equiv f(x)$$

for $0 \leq x < 1$. Thus

$$\lim_{x \to 1^-} f(x) = 0 \,.$$

But

$$\lim_{j \to \infty} \lim_{x \to 1^-} f_j(x) = \lim_{j \to \infty} 1 = 1 \,.$$

Thus the two dual limits in the theorem are unequal in this example. But of course the functions f_j do not converge uniformly. □

Parallel with our notion of Cauchy sequence of numbers, we have a concept of Cauchy sequence of functions in the uniform sense:

Definition 8.12 A sequence of functions f_j on a domain S is called *a uniformly Cauchy sequence* if, for each $\epsilon > 0$, there is an $N > 0$ such that, if $j, k > N$, then

$$|f_j(x) - f_k(x)| < \epsilon \quad \forall x \in S \,.$$

The key point for "uniformly Cauchy" sequence of functions is that the choice of N does not depend on x.

Proposition 8.13 *A sequence of functions f_j is uniformly Cauchy on a domain S if and only if the sequence converges uniformly to a limit function f on the domain S.*

Proof: The proof is straightforward and is assigned as an exercise. □

We will use the last two results in our study of the limits of differentiable functions. First we consider an example.

EXAMPLE 8.14 Define the function

$$f_j(x) = \begin{cases} 0 & \text{if} \quad x \leq 0 \\ jx^2 & \text{if} \quad 0 < x \leq 1/(2j) \\ x - 1/(4j) & \text{if} \quad 1/(2j) < x < \infty \end{cases}$$

We leave it as an exercise for you to check that the functions f_j converge uniformly on the entire real line to the function

$$f(x) = \begin{cases} 0 & \text{if} \quad x \leq 0 \\ x & \text{if} \quad x > 0 \end{cases}$$

(draw a sketch to help you see this). Notice that each of the functions f_j is continuously differentiable on the entire real line, but f is not differentiable at 0. □

It turns out that we must strengthen our convergence hypotheses if we want the limit process to respect differentiation. The basic result is this:

Theorem 8.15 *Suppose that a sequence f_j of differentiable functions on an open interval I converges pointwise to a limit function f. Suppose further that the sequence f_j' converges uniformly on I to a limit function g. Then the limit function f is differentiable on I and $f'(x) = g(x)$ for all $x \in I$.*

Proof: Let $\epsilon > 0$. The sequence $\{f_j'\}$ is uniformly Cauchy. Therefore we may choose N so large that $j, k > N$ implies that

$$\left| f_j'(x) - f_k'(x) \right| < \frac{\epsilon}{2} \quad \forall x \in I. \tag{8.15.1}$$

Fix a point $P \in I$. Define

$$\mu_j(x) = \frac{f_j(x) - f_j(P)}{x - P}$$

for $x \in I, x \neq P$. It is our intention to apply Theorem 8.10 above to the functions μ_j.

First notice that, for each j, we have

$$\lim_{x \to P} \mu_j(x) = f_j'(P).$$

Thus

$$\lim_{j \to \infty} \lim_{x \to P} \mu_j(x) = \lim_{j \to \infty} f_j'(P) = g(P).$$

That calculates the limits in one order.

On the other hand,

$$\lim_{j \to \infty} \mu_j(x) = \frac{f(x) - f(P)}{x - P} \equiv \mu(x)$$

for $x \in I \setminus \{P\}$. If we can show that this convergence is uniform then Theorem 8.10 applies and we may conclude that

$$\lim_{x \to P} \mu(x) = \lim_{j \to \infty} \lim_{x \to P} \mu_j(x) = \lim_{j \to \infty} f_j'(P) = g(P).$$

But this just says that f is differentiable at P and the derivative equals g. That is the desired result.

To verify the uniform convergence of the μ_j, we apply the Mean Value Theorem to the function $f_j - f_k$. For $x \neq P$ we have

$$\begin{aligned} |\mu_j(x) - \mu_k(x)| &= \frac{1}{|x - P|} \cdot |(f_j(x) - f_k(x)) - (f_j(P) - f_k(P))| \\ &= \frac{1}{|x - P|} \cdot |x - P| \cdot |(f_j - f_k)'(\xi)| \\ &= |(f_j - f_k)'(\xi)| \end{aligned}$$

for some ξ between x and P. But line (8.15.1) guarantees that the last line does not exceed $\epsilon/2$. That shows that the μ_j converge uniformly and concludes the proof. $\qquad\square$

Remark 8.16 A little additional effort shows that we need only assume in the theorem that the functions f_j converge at a single point x_0 in the domain. One of the exercises asks you to prove this assertion.

Notice further that, if we make the additional assumption that each of the functions f'_j is continuous, then the proof of the theorem becomes much easier. For then

$$f_j(x) = f_j(x_0) + \int_{x_0}^x f'_j(t)\, dt$$

by the Fundamental Theorem of Calculus. The hypothesis that the f'_j converge uniformly then implies, by Theorem 8.8, that the integrals converge to

$$\int_{x_0}^x g(t)\, dt\,.$$

The hypothesis that the functions f_j converge at x_0 then allows us to conclude that the sequence $f_j(x)$ converges for every x to $f(x)$ and

$$f(x) = f(x_0) + \int_{x_0}^x g(t)\, dt\,.$$

The Fundamental Theorem of Calculus then yields that $f' = g$ as desired.

EXAMPLE 8.17 Consider the sequence of functions $f_j(x) = j^{-1/2}\sin(jx)$. This sequence converges uniformly to the identically zero function $f(x) \equiv 0$. But $f'_j = j^{1/2}\cos(jx)$ does not converge at any point.

We can sum up this result by saying that

$$\lim_{j\to\infty} \frac{d}{dx} f_j(x) \neq \frac{d}{dx} \lim_{j\to\infty} f_j(x)\,. \qquad\qquad\square$$

Exercises

1. Prove that, if a series of continuous functions converges uniformly, then the sum function is also continuous.

2. If a sequence of functions f_j on a domain $S \subseteq \mathbb{R}$ has the property that $f_j \to f$ uniformly on S, then does it follow that $(f_j)^2 \to f^2$ uniformly on S? What simple additional hypothesis will make your answer affirmative?

3. Let f_j be a uniformly convergent sequence of functions on a common domain S. What would be suitable conditions on a function ϕ to guarantee that $\phi \circ f_j$ converges uniformly on S?

4. Prove that a sequence $\{f_j\}$ of functions converges pointwise if and only if the series

$$f_1 + \sum_{j=2}^{\infty} (f_j - f_{j-1})$$

converges pointwise. Prove the same result for uniform convergence.

5. Assume that f_j are continuous functions on the interval $[0,1]$. Suppose that $\lim_{j \to \infty} f_j(x)$ exists for each $x \in [0,1]$ and defines a function f on $[0,1]$. Further suppose that $f_1 \leq f_2 \leq \cdots$. Can you conclude that f is continuous?

6. Let $f : \mathbb{R} \to \mathbb{R}$ be a function. We say that f is *piecewise constant* if the real line can be written as the infinite disjoint union of intervals and f is constant on each of those intervals. Now let φ be a continuous function on $[a, b]$. Show that φ can be uniformly approximated by piecewise constant functions.

7. Refer to Exercise **6** for terminology. Let f be a piecewise constant function. Show that f is the pointwise limit of polynomials.

8. Prove Proposition 8.13. Refer to the parallel result in Chapter 3 for some hints.

9. Prove the assertion made in Remark 8.16 that Theorem 8.15 is still true if the functions f_j are assumed to converge at just one point (and also that the derivatives f_j' converge uniformly).

* **10.** A function is called "piecewise linear" if it is (i) continuous and (ii) its graph consists of finitely many linear segments. Prove that a continuous function on an interval $[a, b]$ is the uniform limit of a sequence of piecewise linear functions.

* **11.** Construct a sequence of continuous functions $f_j(x)$ that has the property that $f_j(q)$ increases monotonically to $+\infty$ for each rational q but such that, at uncountably many irrationals x, $|f_j(x)| \leq 1$ for infinitely many j.

8.3   Series of Functions

Definition 8.18 The formal expression

$$\sum_{j=1}^{\infty} f_j(x) \,,$$

where the f_j are functions on a common domain S, is called a *series of functions*. For $N = 1, 2, 3, \ldots$ the expression

$$S_N(x) = \sum_{j=1}^{N} f_j(x) = f_1(x) + f_2(x) + \cdots + f_N(x)$$

is called the Nth *partial sum* for the series. In case

$$\lim_{N\to\infty} S_N(x)$$

exists and is finite then we say that the series *converges* at x. Otherwise we say that the series *diverges* at x.

Notice that the question of convergence of a series of functions, which should be thought of as an *addition process*, reduces to a question about the *sequence* of partial sums. Sometimes, as in the next example, it is convenient to begin the series at some index other than $j = 1$.

EXAMPLE 8.19 Consider the series

$$\sum_{j=0}^{\infty} x^j \, .$$

This is the geometric series from Proposition 3.15. It converges absolutely for $|x| < 1$ and diverges otherwise.

By the formula for the partial sums of a geometric series,

$$S_N(x) = \frac{1 - x^{N+1}}{1 - x} \, .$$

For $|x| < 1$ we see that

$$S_N(x) \to \frac{1}{1 - x}. \qquad\qquad \square$$

Definition 8.20 Let

$$\sum_{j=1}^{\infty} f_j(x)$$

be a series of functions on a domain S. If the partial sums $S_N(x)$ converge uniformly on S to a limit function $g(x)$ then we say that the series *converges uniformly* on S.

Of course all of our results about uniform convergence of *sequences* of functions translate, via the sequence of partial sums of a series, to results about uniformly convergent series of functions. For example,

(a) If f_j are continuous functions on a domain S and if the series

$$\sum_{j=1}^{\infty} f_j(x)$$

converges uniformly on S to a limit function f then f is also continuous on S.

(b) If f_j are integrable functions on $[a, b]$ and if

$$\sum_{j=1}^{\infty} f_j(x)$$

converges uniformly on $[a, b]$ to a limit function f then f is also integrable on $[a, b]$ and

$$\int_a^b f(x)\, dx = \sum_{j=1}^{\infty} \int_a^b f_j(x)\, dx\,.$$

You will be asked to provide details of these assertions, as well as a statement and proof of a result about derivatives of series, in the exercises. Meanwhile we turn to an elegant test for uniform convergence that is due to Weierstrass.

Theorem 8.21 (The Weierstrass M-Test) *Let $\{f_j\}_{j=1}^{\infty}$ be functions on a common domain S. Assume that each $|f_j|$ is bounded on S by a constant M_j and that*

$$\sum_{j=1}^{\infty} M_j < \infty\,.$$

Then the series

$$\sum_{j=1}^{\infty} f_j \tag{8.21.1}$$

converges uniformly on the set S.

Proof: By hypothesis, the sequence T_N of partial sums of the series $\sum_{j=1}^{\infty} M_j$ is Cauchy. Given $\epsilon > 0$ there is therefore a number K so large that $q > p > K$ implies that

$$\sum_{j=p+1}^{q} M_j = |T_q - T_p| < \epsilon\,.$$

We may conclude that the partial sums S_N of the original series $\sum f_j$ satisfy, for $q > p > K$,

$$|S_q(x) - S_p(x)| = \left| \sum_{j=p+1}^{q} f_j(x) \right|$$

$$\leq \sum_{j=p+1}^{q} |f_j(x)| \leq \sum_{j=p+1}^{q} M_j < \epsilon.$$

Thus the partial sums $S_N(x)$ of the series (8.21.1) are uniformly Cauchy. The series (8.21.1) therefore converges uniformly. $\qquad\square$

EXAMPLE 8.22 Let us consider the series

$$f(x) = \sum_{j=1}^{\infty} 2^{-j} \sin\left(2^j x\right) \ .$$

The sine terms oscillate so erratically that it would be difficult to calculate partial sums for this series. However, noting that the jth summand $f_j(x) = 2^{-j} \sin(2^j x)$ is dominated in absolute value by 2^{-j}, we see that the Weierstrass M-Test applies to this series. We conclude that the series converges uniformly on the entire real line.

By property **(a)** of uniformly convergent series of continuous functions that was noted above, we may conclude that the function f defined by our series is continuous. It is also 2π-periodic: $f(x + 2\pi) = f(x)$ for every x since this assertion is true for each summand. Since the continuous function f restricted to the compact interval $[0, 2\pi]$ is uniformly continuous (Theorem 5.27), we may conclude that f is uniformly continuous on the entire real line.

However, it turns out that f is nowhere differentiable. The proof of this assertion follows lines similar to the treatment of nowhere differentiable functions in Theorem 6.6. The details will be covered in an exercise.   □

Exercises

1. Prove Dini's theorem: If f_j are continuous functions on a compact set K, $f_1(x) \leq f_2(x) \leq \ldots$ for all $x \in K$, and the f_j converge to a continuous function f on K then in fact the f_j converge *uniformly* to f on K.

2. Use the concept of boundedness of a function to show that the functions $\sin x$ and $\cos x$ cannot be polynomials.

3. Prove that, if p is any polynomial, then there is an N large enough that $e^x > |p(x)|$ for $x > N$. Conclude that the function e^x is not a polynomial.

4. Find a way to prove that $\tan x$ and $\ln x$ are not polynomials.

5. Prove that the series

$$\sum_{j=1}^{\infty} \frac{\sin jx}{j}$$

converges uniformly on compact intervals that do not contain odd multiples of $\pi/2$. (**Hint:** Sum by parts and the result will follow.)

6. Suppose that the sequence $f_j(x)$ on the interval $[0, 1]$ satisfies $|f_j(s) - f_j(t)| \leq |s - t|$ for all $s, t \in [0, 1]$. Further assume that the f_j converge pointwise to a limit function f on the interval $[0, 1]$. Does the series converge uniformly?

7. Prove a comparison test for uniform convergence of series: if f_j, g_j are functions and $0 \leq f_j \leq g_j$ and the series $\sum g_j$ converges uniformly then so also does the series $\sum f_j$.

8. Show by giving an example that the converse of the Weierstrass M-Test is false.

9. If f_j are continuous functions on a domain S and if the series

$$\sum_{j=1}^{\infty} f_j(x)$$

converges uniformly on S to a limit function f then f is also continuous on S.

10. Prove that if a series $\sum_{j=1}^{\infty} f_j$ of integrable functions on an interval $[a,b]$ is uniformly convergent on $[a,b]$ then the sum function f is integrable and

$$\int_a^b f(x)\, dx = \sum_{j=1}^{\infty} \int_a^b f_j(x)\, dx\,.$$

* **11.** Give an example of a series of functions on the interval $[0,1]$ that converges pointwise but does not converge uniformly on any subinterval.

12. Formulate and prove a result about the derivative of the sum of a convergent series of differentiable functions.

* **13.** Let $0 < \alpha \le 1$. Prove that the series

$$\sum_{j=1}^{\infty} 2^{-j\alpha} \sin\left(2^j x\right)$$

defines a function f that is nowhere differentiable. To achieve this end, follow the scheme that was used to prove Theorem 6.6: **a)** Fix x; **b)** for h small, choose M such that 2^{-M} is approximately equal to $|h|$; **c)** break the series up into the sum from 1 to $M-1$, the single summand $j = M$, and the sum from $j = M+1$ to ∞. The middle term has very large Newton quotient and the first and last terms are relatively small.

* **14.** Prove that the sequence of functions $f_j(x) = \sin(jx)$ has no subsequence that converges at every x.

8.4 The Weierstrass Approximation Theorem

The name Weierstrass has occurred frequently in this chapter. In fact Karl Weierstrass (1815–1897) revolutionized analysis with his examples and theorems. This section is devoted to one of his most striking results. We introduce it with a motivating discussion.

It is natural to wonder whether the usual functions of calculus—$\sin x$, $\cos x$, and e^x, for instance—are actually polynomials of some very high degree. Since

polynomials are so much easier to understand than these transcendental functions, an affirmative answer to this question would certainly simplify mathematics. Of course a moment's thought shows that this wish is impossible: a polynomial of degree k has at most k real roots. Since sine and cosine have infinitely many real roots they cannot be polynomials. A polynomial of degree k has the property that if it is differentiated enough times (namely, $k+1$ times) then the derivative is zero. Since this is not the case for e^x, we conclude that e^x cannot be a polynomial. The exercises of the last section discuss other means for distinguishing the familiar transcendental functions of calculus from polynomial functions.

In calculus we learned of a formal procedure, called Taylor series, for associating polynomials with a given function f. In some instances these polynomials form a sequence that converges back to the original function. Of course the method of the Taylor expansion has no hope of working unless f is infinitely differentiable. Even then, it turns out that the Taylor series rarely converges back to the original function—see the discussion at the end of Section 9.2. Nevertheless, Taylor's theorem with remainder might cause us to speculate that any reasonable function can be approximated in some fashion by polynomials. In fact the theorem of Weierstrass gives a spectacular affirmation of this speculation:

Theorem 8.23 (Weierstrass Approximation Theorem) *Let f be a continuous function on an interval $[a, b]$. Then there is a sequence of polynomials $p_j(x)$ with the property that the sequence p_j converges uniformly on $[a, b]$ to f.*

In a few moments we shall prove this theorem in detail. Let us first consider some of its consequences. A restatement of the theorem would be that, given a continuous function f on $[a, b]$ and an $\epsilon > 0$, there is a polynomial p such that

$$|f(x) - p(x)| < \epsilon$$

for every $x \in [a, b]$. If one were programming a computer to calculate values of a fairly wild function f, the theorem guarantees that, up to a given degree of accuracy, one could use a polynomial instead (which would in fact be much easier for the computer to handle). Advanced techniques can even tell what degree of polynomial is needed to achieve a given degree of accuracy. The proof that we shall present also suggests how this might be done.

Let f be the Weierstrass nowhere differentiable function. The theorem guarantees that, on any compact interval, f is the uniform limit of polynomials. Thus even the uniform limit of infinitely differentiable functions need not be differentiable—even at one point. This explains why the hypotheses of Theorem 8.15 needed to be so stringent.

We shall break up the proof of the Weierstrass Approximation Theorem into a sequence of lemmas.

Lemma 8.24 *Let ψ_j be a sequence of continuous functions on the interval $[-1, 1]$ with the following properties:*

(i) $\psi_j(x) \geq 0$ for all x;

(ii) $\int_{-1}^{1} \psi_j(x)\, dx = 1$ for each j;

(iii) For any $\delta > 0$ we have

$$\lim_{j \to \infty} \int_{\delta \leq |x| \leq 1} \psi_j(x)\, dx = 0\,.$$

If f is a continuous function on the real line which is identically zero off the interval $[0, 1]$ then the functions

$$f_j(x) = \int_{-1}^{1} \psi_j(t) f(x - t)\, dt$$

converge uniformly on the interval $[0, 1]$ to $f(x)$.

Proof: By multiplying f by a constant we may assume that $\sup |f| = 1$. Let $\epsilon > 0$. Since f is uniformly continuous on the interval $[0, 1]$ we may choose a $\delta > 0$ such that if $x, t \in [0, 1]$ and if $|x - t| < \delta$ then $|f(x) - f(t)| < \epsilon/2$. By property **(iii)** above, we may choose an N so large that $j > N$ implies that $|\int_{\delta \leq |t| \leq 1} \psi_j(t)\, dt| < \epsilon/4$. Then, for any $x \in [0, 1]$, we have

$$
\begin{aligned}
|f_j(x) - f(x)| &= \left| \int_{-1}^{1} \psi_j(t) f(x - t)\, dt - f(x) \right| \\
&= \left| \int_{-1}^{1} \psi_j(t) f(x - t)\, dt - \int_{-1}^{1} \psi_j(t) f(x)\, dt \right|.
\end{aligned}
$$

Notice that, in the last line, we have used fact **(ii)** about the functions ψ_j to multiply the term $f(x)$ by 1 in a clever way. Now we may combine the two integrals to find that the last line

$$
\begin{aligned}
&= \left| \int_{-1}^{1} (f(x - t) - f(x)) \psi_j(t)\, dt \right| \\
&\leq \int_{-\delta}^{\delta} |f(x - t) - f(x)| \psi_j(t)\, dt \\
&\quad + \int_{\delta \leq |t| \leq 1} |f(x - t) - f(x)| \psi_j(t)\, dt \\
&= A + B\,.
\end{aligned}
$$

To estimate term A, we recall that, for $|t| < \delta$, we have $|f(x - t) - f(x)| < \epsilon/2$; hence

$$A \leq \int_{-\delta}^{\delta} \frac{\epsilon}{2} \psi_j(t)\, dt \leq \frac{\epsilon}{2} \cdot \int_{-1}^{1} \psi_j(t)\, dt = \frac{\epsilon}{2}\,.$$

For B we write

$$
\begin{aligned}
B &\leq \int_{\delta \leq |t| \leq 1} 2 \cdot \sup |f| \cdot \psi_j(t)\, dt \\
&\leq 2 \cdot \int_{\delta \leq |t| \leq 1} \psi_j(t)\, dt \\
&< 2 \cdot \frac{\epsilon}{4} = \frac{\epsilon}{2},
\end{aligned}
$$

where in the penultimate line we have used the choice of j. Adding together our estimates for A and B, and noting that these estimates are independent of the choice of x, yields the result. $\qquad\square$

Lemma 8.25 Define $\psi_j(t) = k_j \cdot (1 - t^2)^j$, where the positive constants k_j are chosen so that $\int_{-1}^{1} \psi_j(t)\, dt = 1$. Then the functions ψ_j satisfy the properties **(i)**–**(iii)** of the last lemma.

Proof: Of course property **(ii)** is true by design. Property **(i)** is obvious. In order to verify property **(iii)**, we need to estimate the size of k_j.

Notice that

$$
\begin{aligned}
\int_{-1}^{1} (1 - t^2)^j\, dt &= 2 \cdot \int_0^1 (1 - t^2)^j\, dt \\
&\geq 2 \cdot \int_0^{1/\sqrt{j}} (1 - t^2)^j\, dt \\
&\geq 2 \cdot \int_0^{1/\sqrt{j}} (1 - jt^2)\, dt\,,
\end{aligned}
$$

where we have used the binomial theorem. But this last integral is easily evaluated and equals $4/(3\sqrt{j})$. We conclude that

$$
\int_{-1}^{1} (1 - t^2)^j\, dt > \frac{1}{\sqrt{j}}\,.
$$

As a result, $k_j < \sqrt{j}$.

Now, to verify property **(iii)** of the lemma, we notice that, for $\delta > 0$ fixed and $\delta \leq |t| \leq 1$, it holds that

$$
|\psi_j(t)| \leq k_j \cdot (1 - \delta^2)^j \leq \sqrt{j} \cdot (1 - \delta^2)^j
$$

and this expression tends to 0 as $j \to \infty$. Thus $\psi_j \to 0$ uniformly on $\{t : \delta \leq |t| \leq 1\}$. It follows that the ψ_j satisfy property **(iii)** of the lemma. $\square$

Proof of the Weierstrass Approximation Theorem: We may assume without loss of generality (just by changing coordinates) that f is a continuous function on the interval $[0, 1]$. After adding a linear function (which is a polynomial)

to f, we may assume that $f(0) = f(1) = 0$. Thus f may be continued/extended to be a continuous function which is identically zero on $\mathbb{R} \setminus [0, 1]$.

Let ψ_j be as in Lemma 8.25 and form f_j as in Lemma 8.24. Then we know that the f_j converge uniformly on $[0, 1]$ to f. Finally,

$$
\begin{aligned}
f_j(x) &= \int_{-1}^{1} \psi_j(t) f(x - t)\, dt \\
&= \int_0^1 \psi_j(x - t) f(t)\, dt \\
&= k_j \int_0^1 (1 + (x - t)^2)^j f(t)\, dt\,.
\end{aligned}
$$

But multiplying out the expression $(1 + (x - t)^2)^j$ in the integrand then shows that f_j is a polynomial of degree at most $2j$ in x. Thus we have constructed a sequence of polynomials f_j that converges uniformly to the function f on the interval $[0, 1]$. $\square$

EXAMPLE 8.26 The Weierstrass nowhere differentiable function is a continuous function on $[0, 1]$ that is not differentiable at any point. Nevertheless, it is (by the Weierstrass Approximation Theorem) uniformly approximable by polynomials.

Of course the uniform limit of polynomials will be continuous, so we can only consider continuous functions in this context.

Exercises

1. If f is a continuous function on the interval $[a, b]$ and if

$$
\int_a^b f(x) p(x)\, dx = 0
$$

   for every polynomial p, then prove that f must be the zero function. (**Hint:** Use Weierstrass's Approximation Theorem.)

2. Let $\{f_j\}$ be a sequence of continuous functions on the real line. Suppose that the f_j converge uniformly to a function f. Prove that

$$
\lim_{j \to \infty} f_j(x + 1/j) = f(x)
$$

   uniformly on any bounded interval.

   Can any of these hypotheses be weakened?

3. Prove that the Weierstrass Approximation Theorem fails if we restrict attention to polynomials of degree less than or equal to 1000.

4. Is the Weierstrass Approximation Theorem true if we restrict ourselves to only using polynomials of even degree?

5. Is the Weierstrass Approximation Theorem true if we restrict ourselves to only using polynomials with coefficients of size not exceeding 1?

6. Use the polar form of complex numbers (that is, $z = re^{i\theta}$) to show that, on the unit circle, trigonometric polynomials and ordinary polynomials are really the same thing.

7. The Weierstrass approximation theorem says that, if f is a continuous function on $[0, 1]$, then there is a sequence of polynomials p_j that converges uniformly on $[0, 1]$ to f. Now take f to be continuously differentiable. The Weierstrass theorem applies to give a sequence p_j that converges to f. What can you say about p_j' converging to f'?

*   8. Use the Weierstrass Approximation Theorem and Mathematical Induction to prove that, if f is k times continuously differentiable on an interval $[a, b]$, then there is a sequence of polynomials p_j with the property that

$$p_j \to f$$

uniformly on $[a, b]$,

$$p_j' \to f'$$

uniformly on $[a, b]$,

$$\cdots$$

$$p_j^{(k)} \to f^{(k)}$$

uniformly on $[a, b]$.

*   9. Let $a < b$ be real numbers. Call a function of the form

$$f(x) = \begin{cases} 1 & \text{if} \quad a \le x \le b \\ 0 & \text{if} \quad x < a \text{ or } x > b \end{cases}$$

a *characteristic function* for the interval $[a, b]$. Then a function of the form

$$g(x) = \sum_{j=1}^{k} a_j \cdot f_j(x),$$

with the f_j characteristic functions of intervals $[a_j, b_j]$, is called *simple*. Prove that any continuous function on an interval $[c, d]$ is the uniform limit of a sequence of simple functions. (**Hint:** The proof of this assertion is conceptually simple; do *not* imitate the proof of the Weierstrass Approximation Theorem.)

* **10.** Define a *trigonometric polynomial* to be a function of the form

$$\sum_{j=1}^{k} a_j \cdot \cos jx + \sum_{j=1}^{\ell} b_j \cdot \sin jx \, .$$

Prove a version of the Weierstrass Approximation Theorem on the interval $[0, 2\pi]$ for 2π-periodic continuous functions and with the phrase "trigonometric polynomial" replacing "polynomial." (**Hint:** Prove that

$$\sum_{\ell=-j}^{j} \left(1 - \frac{|\ell|}{j+1}\right)(\cos \ell t) =$$

$$\frac{1}{j+1}\left(\frac{\sin \frac{j+1}{2}t}{\sin \frac{1}{2}t}\right)^2 \, .$$

Use these functions as the ψ_js in the proof of Weierstrass's theorem.)

* **11.** There is a version of the Weierstrass Approximation Theorem on the unit square $[0, 1] \times [0, 1] \subseteq \mathbb{R}^2$. What should it say?

Chapter 9

Elementary Transcendental Functions

9.1   Power Series

A series of the form

$$\sum_{j=0}^{\infty} a_j (x - c)^j$$

is called a *power series* expanded about the point c. Our first task is to determine the nature of the set on which a power series converges.

Proposition 9.1 *Assume that the power series*

$$\sum_{j=0}^{\infty} a_j (x - c)^j$$

converges at the value $x = d$ *with* $d \neq c$. *Let* $r = |d - c|$. *Then the series converges uniformly and absolutely on compact subsets of* $\mathcal{I} = \{x : |x - c| < r\}$.

Proof: We may take the compact subset of $\mathcal{I}$ to be $K = [c - s, c + s]$ for some number $0 < s < r$. For $x \in K$ it then holds that

$$\sum_{j=0}^{\infty} \left| a_j (x - c)^j \right| = \sum_{j=0}^{\infty} \left| a_j (d - c)^j \right| \cdot \left| \frac{x - c}{d - c} \right|^j .$$

In the sum on the right, the first expression in absolute values is bounded by some constant C (by the convergence hypothesis). The quotient in absolute values is majorized by $L = s/r < 1$. The series on the right is thus dominated by

$$\sum_{j=0}^{\infty} C \cdot L^j .$$

This geometric series converges. By the Weierstrass M-Test, the original series converges absolutely and uniformly on K. □

An immediate consequence of the proposition is that the set on which the power series

$$\sum_{j=0}^{\infty} a_j (x - c)^j$$

converges is an interval centered about c. We call this set the *interval of convergence*. The series will converge absolutely and uniformly on compact subsets of the interval of convergence. The *radius* of the interval of convergence (called the *radius of convergence*) is defined to be half its length. Whether convergence holds at the endpoints of the interval will depend on the particular series being studied. Ad hoc methods must be used to check the endpoints. Let us use the notation $\mathcal{C}$ to denote the *open interval of convergence*.

It happens that, if a power series converges at either of the endpoints of its interval of convergence, then the convergence is uniform up to that endpoint. This is a consequence of Abel's partial summation test; details will be explored in the exercises.

EXAMPLE 9.2 Consider the power series

$$\sum_{j=1}^{\infty} 2^j x^j .$$

We may apply the Root Test to this series to see that

$$|a_j|^{1/j} = |2^j x^j|^{1/j} = 2|x| .$$

This expression is less than 1 precisely when $|x| < 1/2$. Thus the open interval of convergence for this power series is $(-1/2, 1/2)$. We can easily check by hand that the series does *not* converge at the endpoints. □

On the interval of convergence $\mathcal{C}$, the power series defines a function f. Such a function is said to be *real analytic*. More precisely, we have

Definition 9.3 A function f, with domain an open set $U \subseteq \mathbb{R}$ and range either the real or the complex numbers, is called *real analytic* if, for each $c \in U$, the function f may be represented by a convergent power series on an interval of positive radius centered at c:

$$f(x) = \sum_{j=0}^{\infty} a_j (x - c)^j .$$

EXAMPLE 9.4 The function $f(x) = 1/(1 + x^2)$ is real analytic on the interval $(-1, 1)$. This is true because

$$\frac{1}{1 + x^2} = \sum_{j=0}^{\infty} (-x^2)^j .$$

In actuality, f is real analytic on the entire real line. But it requires power series centered at points other than the origin to see this. The entire matter is best explained in the context of complex variables, and this point of view is explained below. $\square$

We need to know both the algebraic and the calculus properties of a real analytic function: is it continuous? differentiable? How does one add/subtract/ multipy/divide two such functions?

Proposition 9.5 *Let*

$$\sum_{j=0}^{\infty} a_j(x-c)^j \quad \text{and} \quad \sum_{j=0}^{\infty} b_j(x-c)^j$$

be two power series with intervals of convergence $\mathcal{C}_1$ and $\mathcal{C}_2$ centered at c. Let $f_1(x)$ be the function defined by the first series on $\mathcal{C}_1$ and $f_2(x)$ the function defined by the second series on $\mathcal{C}_2$. Then, on their common domain $\mathcal{C} = \mathcal{C}_1 \cap \mathcal{C}_1$, it holds that

(1) $f(x) \pm g(x) = \sum_{j=0}^{\infty}(a_j \pm b_j)(x-c)^j$;

(2) $f(x) \cdot g(x) = \sum_{m=0}^{\infty} \sum_{j+k=m}(a_j \cdot b_k)(x-c)^m$.

Proof: Let

$$A_N = \sum_{j=0}^{N} a_j(x-c)^j \quad \text{and} \quad B_N = \sum_{j=0}^{N} b_j(x-c)^j$$

be, respectively, the Nth partial sums of the power series that define f and g. If C_N is the Nth partial sum of the series

$$\sum_{j=0}^{\infty}(a_j \pm b_j)(x-c)^j$$

then

$$
\begin{aligned}
f(x) \pm g(x) &= \lim_{N \to \infty} A_N \pm \lim_{N \to \infty} B_N = \lim_{N \to \infty} [A_N \pm B_N] \\
&= \lim_{N \to \infty} C_N = \sum_{j=0}^{\infty}(a_j \pm b_j)(x-c)^j \, .
\end{aligned}
$$

This proves **(1)**.
    For **(2)**, let

$$D_N = \sum_{m=0}^{N} \sum_{j+k=m}(a_j \cdot b_k)(x-c)^m \quad \text{and} \quad R_N = \sum_{j=N+1}^{\infty} b_j(x-c)^j \, .$$

We have

$$
\begin{aligned}
D_N &= a_0 B_N + a_1(x-c)B_{N-1} + \cdots + a_N(x-c)^N B_0 \\
&= a_0(g(x) - R_N) + a_1(x-c)(g(x) - R_{N-1}) \\
&\quad + \cdots + a_N(x-c)^N(g(x) - R_0) \\
&= g(x)\sum_{j=0}^{N} a_j(x-c)^j \\
&\quad - [a_0 R_N + a_1(x-c)R_{N-1} + \cdots + a_N(x-c)^N R_0].
\end{aligned}
$$

Clearly,

$$
g(x)\sum_{j=0}^{N} a_j(x-c)^j
$$

converges to $g(x)f(x)$ as N approaches ∞. In order to show that $D_N \to g \cdot f$, it will thus suffice to show that

$$
\left| a_0 R_N + a_1(x-c)R_{N-1} + \cdots + a_N(x-c)^N R_0 \right|
$$

converges to 0 as N approaches ∞. Fix x. Now we know that

$$
\sum_{j=0}^{\infty} a_j(x-c)^j
$$

is absolutely convergent so we may set

$$
A = \sum_{j=0}^{\infty} |a_j||x-c|^j.
$$

Also $\sum_{j=0}^{\infty} b_j(x-c)^j$ is convergent. Therefore, given $\epsilon > 0$, we can find N_0 so that $N > N_0$ implies $|R_N| < \epsilon$. Thus we have

$$
\begin{aligned}
&\left| a_0 R_N + a_1(x-c)R_{N-1} + \cdots + a_N(x-c)^N R_0 \right| \\
&\leq \left| a_0 R_N + \cdots + a_{N-N_0}(x-c)^{N-N_0} R_{N_0} \right| \\
&\quad + \left| a_{N-N_0+1}(x-c)^{N-N_0+1} R_{N_0-1} + \cdots + a_N(x-c)^N R_0 \right| \\
&\leq \sup_{M \geq N_0} R_M \cdot \left(\sum_{j=0}^{\infty} |a_j||x-c|^j \right) \\
&\quad + \left| a_{N-N_0+1}(x-c)^{N-N_0+1} R_{N_0-1} \cdots + a_N(x-c)^N R_0 \right| \\
&\leq \epsilon \cdot A + \left| a_{N-N_0+1}(x-c)^{N-N_0+1} R_{N_0-1} \cdots + a_N(x-c)^N R_0 \right|.
\end{aligned}
$$

Thus

$$
\begin{aligned}
&\left| a_0 R_N + a_1(x-c)R_{N-1} + \cdots + a_N(x-c)^N R_0 \right| \\
&\leq \epsilon \cdot A + M \cdot \sum_{j=N-N_0+1}^{N} |a_j||x-c|^j,
\end{aligned}
$$

where M is an upper bound for $|R_j(x)|$. Since the series defining A converges, we find on letting $N \to \infty$ that

$$\limsup_{N\to\infty} \left| a_0 R_N + a_1(x-c)R_{N-1} + \cdots + a_N(x-c)^N R_0 \right| \le \epsilon \cdot A.$$

Since $\epsilon > 0$ was arbitrary, we may conclude that

$$\lim_{N\to\infty} \left| a_0 R_N + a_1(x-c)R_{N-1} + \cdots + a_N(x-c)^N R_0 \right| = 0. \qquad \square$$

Remark 9.6 Observe that the form of the product of two power series provides some motivation for the form that the product of numerical series took in Theorem 3.49.

Next we turn to division of real analytic functions. If f and g are real analytic functions, both defined on an open interval I, and if g does not vanish on I, then we would like f/g to be a well-defined real analytic function (it surely is a well-defined *function*) and we would like to be able to calculate its power series expansion by formal long division. This is what the next result tells us.

Proposition 9.7 *Let f and g be real analytic functions, both of which are defined on an open interval I. Assume that g does not vanish on I. Then the function*

$$h(x) = \frac{f(x)}{g(x)}$$

is real analytic on I. Moreover, if I is centered at the point c and if

$$f(x) = \sum_{j=0}^{\infty} a_j(x-c)^j \quad \text{and} \quad g(x) = \sum_{j=0}^{\infty} b_j(x-c)^j,$$

then the power series expansion of h about c may be obtained by formal long division of the latter series into the former. That is, the zeroeth coefficient c_0 of h is

$$c_0 = a_0/b_0,$$

the order one coefficient c_1 is

$$c_1 = \frac{1}{b_0}\left(a_1 - \frac{a_0 b_1}{b_0} \right),$$

etc.

Proof: If we can show that the power series

$$\sum_{j=0}^{\infty} c_j(x-c)^j$$

converges on I then the result on multiplication of series in Proposition 9.5 yields this new result. There is no loss of generality in assuming that $c = 0$. Assume for the moment that $b_1 \neq 0$.

Notice that one may check inductively that, for $j \geq 1$,

$$c_j = \frac{1}{b_0} \left(a_j - b_1 \cdot c_{j-1} \right) . \tag{9.7.1}$$

Without loss of generality, we may scale the a_js and the b_js and assume that the radius of I is $1 + \epsilon$, some $\epsilon > 0$. Then we see from (9.7.1) that

$$|c_j| \leq C \cdot (|a_j| + |c_{j-1}|) ,$$

where $C = \max\{|1/b_0|, |b_1/b_0|\}$. It follows that

$$|c_j| \leq C' \cdot (1 + |a_j| + |a_{j-1}| + \cdots + |a_0|) ,$$

Since the radius of I exceeds 1, $\sum |a_j| < \infty$ and we see that the $|c_j|$ are bounded. Hence the power series with coefficients c_j has radius of convergence 1.

In case $b_1 = 0$ then the role of b_1 is played by the first nonvanishing $b_m, m > 1$. Then a new version of formula (9.7.1) is obtained and the argument proceeds as before. □

EXAMPLE 9.8 In practice it is often useful to calculate f/g by expanding g in a "geometric series." To illustrate this idea, we assume for simplicity that f and g are real analytic in a neighborhood of 0. Then

$$
\begin{aligned}
\frac{f(x)}{g(x)} &= f(x) \cdot \frac{1}{g(x)} \\
&= f(x) \cdot \frac{1}{b_0 + b_1 x + \cdots} \\
&= f(x) \cdot \frac{1}{b_0} \cdot \frac{1}{1 + (b_1/b_0)x + \cdots} .
\end{aligned}
$$

Now we use the fact that, for β small,

$$\frac{1}{1 - \beta} = 1 + \beta + \beta^2 + \cdots .$$

Setting $\beta = -(b_1/b_0)x - (b_2/b_0)x^2 - \cdots$, we thus find that

$$\frac{f(x)}{g(x)} = \frac{f(x)}{b_0} \cdot \left(1 + [-(b_1/b_0)x - (b_2/b_0)x^2 - \cdots] + [-(b_1/b_0)x - (b_2/b_0)x^2 - \cdots]^2 + \cdots \right) .$$

We explore this technique further in the exercises. □

Exercises

1. Prove that the composition of two real analytic functions, when the composition makes sense, is also real analytic.

2. Prove that
$$\sin^2 x + \cos^2 x = 1$$
   directly from the power series expansions.

3. Verify the formula
$$\frac{1}{1 - \beta} = 1 + \beta + \beta^2 + \cdots$$
   for $|\beta| < 1$.

4. Use the technique described at the end of this section to calculate the first five terms of the power series expansion of $\sin x / e^x$ about the origin.

5. Show that the solution of the differential equation $y' + y = x$ will be real analytic.

6. Provide the details of the method for dividing real analytic functions that is described in Example 9.8.

* 7. Let $f(x) = \sum_{j=0}^{\infty} a_j x^j$ be defined by a power series convergent on the interval $(-r, r)$ and let Z denote those points in the interval where f vanishes. Prove that if Z has an accumulation point in the interval then $f \equiv 0$. (**Hint:** If a is the accumulation point, expand f in a power series about a. What is the first nonvanishing term in that expansion?)

* 8. Verify that the function
$$f(x) = \begin{cases} 0 & \text{if} \quad x = 0 \\ e^{-1/x^2} & \text{if} \quad x \neq 0 \end{cases}$$
   is infinitely differentiable on all of $\mathbb{R}$ and that $f^{(k)}(0) = 0$ for every k. However, f is not real analytic.

* 9. Prove the assertion from the text that, if a power series converges at an endpoint of the interval of convergence, then the convergence is uniform up to that endpoint.

9.2   More on Power Series: Convergence Issues

We now introduce the *Hadamard formula* for the radius of convergence of a power series.

Lemma 9.9 (Hadamard) *For the power series*

$$\sum_{j=0}^{\infty} a_j (x - c)^j \,,$$

define A and ρ by

$$A = \limsup_{n \to \infty} |a_n|^{1/n} \,,$$

$$\rho = \begin{cases} 0 & \text{if } A = \infty, \\ 1/A & \text{if } 0 < A < \infty, \\ \infty & \text{if } A = 0 \,, \end{cases}$$

then ρ is the radius of convergence of the power series about c.

Proof: Observing that

$$\limsup_{n \to \infty} |a_n (x - c)^n|^{1/n} = A|x - c| \,,$$

we see that the lemma is an immediate consequence of the Root Test. □

EXAMPLE 9.10 Consider the power series

$$\sum_{j=1}^{\infty} j x^j \,.$$

We calculate that

$$A = \limsup_{n \to \infty} |a_n|^{1/n} = \limsup_{n \to \infty} n^{1/n} = 1 \,.$$

It follows that the radius of convergence of the power series is $1/1 = 1$. So the open interval of convergence is $\mathcal{C} = (-1, 1)$. The series does *not* converge at the endpoints. □

Corollary 9.11 *The power series*

$$\sum_{j=0}^{\infty} a_j (x - c)^j$$

has radius of convergence ρ if and only if, when $0 < R < \rho$, there exists a constant $0 < C = C_R$ such that

$$|a_j| \le \frac{C}{R^j} \,.$$

EXAMPLE 9.12 The series

$$\sum_{j=0}^{\infty} \frac{3^j}{j^2+1} x^j$$

satisfies

$$|a_j| \leq 3^j .$$

It follows from the corollary then that the radius of convergence of the series is $1/3$. $\square$

From the power series

$$\sum_{j=0}^{\infty} a_j (x-c)^j$$

it is natural to create the *derived series*

$$\sum_{j=1}^{\infty} j a_j (x-c)^j$$

using term-by-term differentiation.

Proposition 9.13 *The radius of convergence of the derived series is the same as the radius of convergence of the original power series.*

Proof: We observe that

$$
\begin{aligned}
\limsup_{j\to\infty} |j a_j|^{1/j} &= \lim_{j\to\infty} j^{-1/j} \limsup_{j\to\infty} |j a_j|^{1/j} \\
&= \limsup_{j\to\infty} |a_j|^{1/j} .
\end{aligned}
$$

So the result follows from the Hadamard formula. $\square$

Proposition 9.14 *Let f be a real analytic function defined on an open interval I. Then f is continuous and has continuous, real analytic derivatives of all orders. In fact the derivatives of f are obtained by differentiating its series representation term by term.*

Proof: Since, for each $c \in I$, the function f may be represented by a convergent power series about c with positive radius of convergence, we see that, in a sufficiently small open interval about each $c \in I$, the function f is the uniform limit of a sequence of continuous functions: the partial sums of the power series representing f. It follows that f is continuous at c. Since the radius of convergence of the derived series is the same as that of the original series, it also follows that the derivatives of the partial sums converge uniformly on an open interval about c to a continuous function. It then follows from Theorem 8.15 that f is differentiable and its derivative is the function defined by the derived series. By induction, f has continuous derivatives of all orders at c. $\square$

EXAMPLE 9.15 The series

$$\sum_{j=0}^{\infty} x^j$$

has derived series

$$\sum_{j=0}^{\infty} j x^{j-1} \,.$$

Of course the original series converges to $1/(1-x)$ and the derived series converges to $1/(1-x)^2$. □

We can now show that a real analytic function has a unique power series representation at any point.

Corollary 9.16 *If the function f is represented by a convergent power series on an interval of positive radius centered at c,*

$$f(x) = \sum_{j=0}^{\infty} a_j (x-c)^j \,,$$

then the coefficients of the power series are related to the derivatives of the function by

$$a_n = \frac{f^{(n)}(c)}{n!} \,.$$

Proof: This follows readily by differentiating both sides of the above equation n times, as we may by the proposition, and evaluating at $x = c$. □

EXAMPLE 9.17 The function

$$f(x) = x \sin x$$

has power series expansion about 0 with coefficients

$$a_3 = \frac{1}{3!} \frac{d^3}{dx^3} f(x) \bigg|_{x=0} = 0$$

and

$$a_4 = \frac{1}{4!} \frac{d^4}{dx^4} f(x) \bigg|_{x=0} = -\frac{1}{3!} \,.$$ □

Finally, we note that integration of power series is as well-behaved as differentiation.

Proposition 9.18 *The power series*

$$\sum_{j=0}^{\infty} a_j (x-c)^j$$

and the series

$$\sum_{j=0}^{\infty} \frac{a_j}{j+1}(x-c)^{j+1}$$

obtained from term-by-term integration have the same radius of convergence, and the function F defined by

$$F(x) = \sum_{j=0}^{\infty} \frac{a_j}{j+1}(x-c)^{j+1}$$

on the common interval of convergence satisfies

$$F'(x) = \sum_{j=0}^{\infty} a_j(x-c)^j = f(x).$$

Proof: The proof is left to the exercises.   □

It is sometimes convenient to allow the variable in a power series to be a complex number. In this case we write

$$\sum_{j=0}^{\infty} a_j(z-c)^j\,,$$

where z is the complex argument. We now allow c and the a_js to be complex numbers as well. Noting that the elementary facts about series hold for complex series as well as real series (you should check this for yourself), we see that the arguments of this section show that the domain of convergence of a complex power series is a *disc* in the complex plane with radius ρ given as follows:

$$A = \limsup_{n\to\infty} |a_n|^{1/n}$$

$$\rho = \begin{cases} 0 & \text{if} \quad A = \infty \\ 1/A & \text{if} \quad 0 < A < \infty \\ \infty & \text{if} \quad A = 0\,. \end{cases}$$

The proofs in this section apply to show that convergent complex power series may be added, subtracted, multiplied, and divided (provided that we do not divide by zero) on their common domains of convergence. They may also be differentiated and integrated term by term.

These observations about complex power series will be useful in the next section.

EXAMPLE 9.19 The function $f(x) = 1/(1 + x^2)$ has power series expansion about the origin given by

$$\sum_{j=0}^{\infty} (-x^2)^j\,.$$

The radius of convergence of the power series is 1, and one might not have anticipated this fact by examining the formula for f.

But if instead one replaces x by z and examines the complex version of the function then one has

$$\widetilde{f}(z) = \frac{1}{1+z^2},$$

and one sees that this function has a singularity at $z = i$. That explains why the radius of convergence is 1. The power series about 0 cannot make sense at the singular point. $\qquad\square$

We conclude this section with a consideration of Taylor series:

Theorem 9.20 (Taylor's Expansion) *For k a nonnegative integer, let f be a $k + 1$ times continuously differentiable function on an open interval $I = (a - \epsilon, a + \epsilon)$. Then, for $x \in I$,*

$$f(x) = \sum_{j=0}^{k} f^{(j)}(a)\frac{(x-a)^j}{j!} + R_{k,a}(x),$$

where

$$R_{k,a}(x) = \int_{a}^{x} f^{(k+1)}(t)\frac{(x-t)^k}{k!}\,dt\,.$$

Proof: We apply integration by parts to the Fundamental Theorem of Calculus to obtain

$$
\begin{aligned}
f(x) &= f(a) + \int_{a}^{x} f'(t)\,dt \\
&= f(a) + \left(f'(t)\frac{(t-x)}{1!}\right)\Big|_{a}^{x} - \int_{a}^{x} f''(t)\frac{(t-x)}{1!}\,dt \\
&= f(a) + f'(a)\frac{(x-a)}{1!} + \int_{a}^{x} f''(t)\frac{x-t}{1!}\,dt\,.
\end{aligned}
$$

Notice that, when we performed the integration by parts, we used $t - x$ as an antiderivative for dt. This is of course legitimate, as a glance at the integration by parts theorem reveals. We have proved the theorem for the case $k = 1$. The result for higher values of k is obtained inductively by repeated applications of integration by parts. $\qquad\square$

Taylor's theorem allows us to associate with any infinitely differentiable function a formal expansion of the form

$$\sum_{j=0}^{\infty} a_j(x-a)^j\,.$$

However, there is no guarantee that this series will converge; even if it does converge, it may not converge back to $f(x)$.

EXAMPLE 9.21 An important example to keep in mind is the function

$$h(x) = \begin{cases} 0 & \text{if} \quad x = 0 \\ e^{-1/x^2} & \text{if} \quad x \neq 0. \end{cases}$$

This function is infinitely differentiable at every point of the real line (including the point 0—use l'Hôpital's Rule). However, all of its derivatives at $x = 0$ are equal to zero (this matter will be treated in the exercises). Therefore the formal Taylor series expansion of h about $a = 0$ is

$$\sum_{j=0}^{\infty} 0 \cdot (x - 0)^j = 0.$$

We see that the formal Taylor series expansion for h converges to the zero function at every x, but not to the original function h itself. $\quad\square$

In fact the theorem tells us that the Taylor expansion of a function f converges to f at a point x if and only if $R_{k,a}(x) \to 0$. In the exercises we shall explore the following more quantitative assertion:

An infinitely differentiable function f on an interval I has Taylor series expansion about $a \in I$ that converges back to f on a neighborhood J of a if and only if there are positive constants C, R such that, for every $x \in J$ and every k, it holds that

$$\left| f^{(k)}(x) \right| \leq C \cdot \frac{k!}{R^k}.$$

The function h in Example 9.21 should not be thought of as an isolated exception. For instance, we know from calculus that the function $f(x) = \sin x$ has Taylor expansion that converges to f at every x. But then, for ϵ small, the function $g_\epsilon(x) = f(x) + \epsilon \cdot h(x)$ has Taylor series that does *not* converge back to $g_\epsilon(x)$ for $x \neq 0$. Similar examples may be generated by using other real analytic functions in place of sine.

Exercises

1. Let f be an infinitely differentiable function on an interval I. If $a \in I$ and there are positive constants C, R such that, for every x in a neighborhood of a and every k, it holds that

$$\left| f^{(k)}(x) \right| \leq C \cdot \frac{k!}{R^k},$$

then prove that the Taylor series of f about a converges to $f(x)$. (**Hint:** estimate the error term.) What is the radius of convergence?

2. Let f be an infinitely differentiable function on an open interval I centered at a. Assume that the Taylor expansion of f about a converges to f at every point of I. Prove that there are constants C, R and a (possibly smaller) interval J centered at a such that, for each $x \in J$, it holds that

$$\left| f^{(k)}(x) \right| \leq C \cdot \frac{k!}{R^k} .$$

3. Give examples of power series, centered at 0, on the interval $(-1, 1)$, which **(a)** converge only on $(-1, 1)$, **(b)** converge only on $[-1, 1)$, **(c)** converge only on $(-1, 1]$, **(d)** converge only on $[-1, 1]$.

4. We know from the text that the real analytic function $1/(1 + x^2)$ is well defined on the entire real line. Yet its power series about 0 only converges on an interval of radius 1.

   How do matters differ for the function $1/(1 - x^2)$?

5. Prove Proposition 9.18.

* 6. The function defined by a power series may extend continuously to an endpoint of the interval of convergence without the series converging at that endpoint. Give an example.

* 7. Prove that, if a function on an interval I has derivatives of all orders which are positive at every point of I, then f is real analytic on I.

* 8. What can you say about the set of convergence of a power series of two real variables?

* 9. Show that the function

$$h(x) = \begin{cases} 0 & \text{if} \quad x = 0 \\ e^{-1/x^2} & \text{if} \quad x \neq 0. \end{cases}$$

   is infinitely differentiable on the entire real line, but it is not real analytic.

* 10. For which x, y does the two-variable power series

$$\sum_j 2^j x^j y^j$$

   converge?

9.3   The Exponential and Trigonometric Functions

We begin by defining the exponential function:

Definition 9.22 The power series

$$\sum_{j=0}^{\infty} \frac{z^j}{j!}$$

converges, by the Ratio Test, for every complex value of z. The function defined thereby is called the *exponential function* and is written $\exp(z)$.

Proposition 9.23 *The function* $\exp(z)$ *satisfies*

$$\exp(a + b) = \exp(a) \cdot \exp(b)$$

for any complex numbers a *and* b.

Proof: We write the right-hand side as

$$\left(\sum_{j=0}^{\infty} \frac{a^j}{j!} \right) \cdot \left(\sum_{j=0}^{\infty} \frac{b^j}{j!} \right) .$$

Now convergent power series may be multiplied term by term. We find that the last line equals

$$\sum_{j=0}^{\infty} \left(\sum_{\ell=0}^{j} \frac{a^{j-\ell}}{(j-\ell)!} \cdot \frac{b^\ell}{\ell!} \right) . \tag{9.23.1}$$

However, the inner sum on the right side of this equation may be written as

$$\frac{1}{j!} \sum_{\ell=0}^{j} \frac{j!}{\ell!(j-\ell)!} a^{j-\ell} b^\ell = \frac{1}{j!} (a+b)^j .$$

It follows that line (9.23.1) equals $\exp(a + b)$. $\qquad \square$

EXAMPLE 9.24 We set $e = \exp(1)$. This is consistent with our earlier treatment of the number e in Section 3.4. The proposition tells us that, for any positive integer k, we have

$$e^k = e \cdot e \cdots e = \exp(1) \cdot \exp(1) \cdots \exp(1) = \exp(k) .$$

If m is another positive integer then

$$(\exp(k/m))^m = \exp(k) = e^k ,$$

whence

$$\exp(k/m) = e^{k/m} .$$

We may extend this formula to *negative* rational exponents by using the fact that $\exp(a) \cdot \exp(-a) = 1$. Thus, for any rational number q,

$$\exp(q) = e^q . \qquad \square$$

Now note that the function exp is increasing and continuous. It follows (this fact is treated in the exercises) that if we set, for any $r \in \mathbb{R}$,

$$e^r = \sup\{q \in \mathbb{Q} : q < r\}$$

(this is a *definition* of the expression e^r) then $e^x = \exp(x)$ for every real x. [You may find it useful to review the discussion of exponentiation in Sections 2.4, 3.4; the presentation here parallels those treatments.] We will adhere to custom and write e^x instead of $\exp(x)$ when the argument of the function is real.

Proposition 9.25 *The exponential function e^x, for $x \in \mathbb{R}$, satisfies*

(a) $e^x > 0$ *for all x;*

(b) $e^0 = 1$;

(c) $(e^x)' = e^x$;

(d) e^x *is strictly increasing;*

(e) *the graph of e^x is asymptotic to the negative x-axis;*

(f) *for each integer $N > 0$ there is a number c_N such that $e^x > c_N \cdot x^N$ when $x > 0$.*

Proof: The first three statements are obvious from the power series expansion for the exponential function.

If $s < t$ then the Mean Value Theorem tells us that there is a number ξ between s and t such that

$$e^t - e^s = (t - s) \cdot e^\xi > 0;$$

hence the exponential function is strictly increasing.

By inspecting the power series we see that $e^x > 1 + x$ hence e^x increases to $+\infty$. Since $e^x \cdot e^{-x} = 1$ we conclude that e^{-x} tends to 0 as $x \to +\infty$. Thus the graph of the exponential function is asymptotic to the negative x-axis.

Finally, by inspecting the power series for e^x, we see that the last assertion is true with $c_N = 1/N!$. □

EXAMPLE 9.26 Let us think about 9.25(c). Which functions satisfy $y' = y$? We may rewrite this equation as

$$\frac{y'}{y} = 1 \, .$$

Now integrate both sides to obtain

$$\ln|y| = x + C \, .$$

Exponentiation now yields

$$|y| = e^C \cdot e^x \,.$$

If we assume that y is a positive function then we can erase the absolute value signs on the left-hand side. And we can rename the constant e^C with the simpler name K. So our equation is

$$y = Ke^x \,.$$

We have discovered that, up to a constant factor, the exponential function is the only function that satisfies 9.25(c).   □

Now we turn to the trigonometric functions. The definition of the trigonometric functions that is found in calculus texts is unsatisfactory because it relies too heavily on a picture and because the continual need to subtract off superfluous multiples of 2π is clumsy. We have nevertheless used the trigonometric functions in earlier chapters to illustrate various concepts. It is time now to give a rigorous definition of the trigonometric functions that is independent of these earlier considerations.

Definition 9.27 The power series

$$\sum_{j=0}^{\infty} (-1)^j \frac{x^{2j+1}}{(2j+1)!}$$

converges at every point of the real line (by the Ratio Test). The function that it defines is called the *sine* function and is usually written $\sin x$.

The power series

$$\sum_{j=0}^{\infty} (-1)^j \frac{x^{2j}}{(2j)!}$$

converges at every point of the real line (by the Ratio Test). The function that it defines is called the *cosine* function and is usually written $\cos x$.

EXAMPLE 9.28 One advantage of having sine and cosine defined with power series is that we can actually use the series to obtain approximate numerical values for these functions. For instance, if we want to know the value of $\sin 1$, we may write

$$\sin 1 \approx 1 - \frac{1^3}{3!} + \frac{1^5}{5!} = 1 - \frac{1}{6} + \frac{1}{120} = \frac{101}{120} \approx 0.84167 \,.$$

The true value of $\sin 1$, determined with a calculator, is 0.84147. So this is a fairly good result. More accuracy can of course be obtained by using more terms of the series.   □

You may recall that the power series that we use to define the sine and cosine functions are precisely the Taylor series expansions for the functions sine

and cosine that were derived in your calculus text. But now we *begin* with the power series and must derive the properties of sine and cosine that we need *from these series*.

In fact the most convenient way to achieve this goal is to proceed by way of the exponential function. [The point here is mainly one of convenience. It can be verified by direct manipulation of the power series that $\sin^2 x + \cos^2 x = 1$ and so forth but the algebra is extremely unpleasant.] The formula in the next proposition is usually credited to Euler.

Proposition 9.29 *The exponential function and the functions* sine *and* cosine *are related by the formula (for x and y real and $i^2 = -1$)*

$$\exp(x + iy) = e^x \cdot (\cos y + i \sin y) \ .$$

Proof: We shall verify the case $x = 0$ and leave the general case for the reader.

Thus we are to prove that

$$e^{iy} = \cos y + i \sin y \ . \tag{9.29.1}$$

Writing out the power series for the exponential, we find that the left-hand side of (9.29.1) is

$$\sum_{j=0}^{\infty} \frac{(iy)^j}{j!}$$

and this equals

$$\left[1 - \frac{y^2}{2!} + \frac{y^4}{4!} - + \cdots \right] + i \left[\frac{y}{1!} - \frac{y^3}{3!} + \frac{y^5}{5!} - + \cdots \right] .$$

Of course the two series on the right are the familiar power series for cosine and sine as specified in Definition 9.27. Thus

$$e^{iy} = \cos y + i \sin y \ ,$$

as desired.   □

EXAMPLE 9.30 We may calculate that

$$e^{i\pi/3} = \cos \frac{\pi}{3} + i \sin \frac{\pi}{3} = \frac{1}{2} + i \frac{\sqrt{3}}{2} \ . \qquad\qquad □$$

In what follows, we think of the formula (9.29.1) as *defining* what we mean by e^{iy}. As a result,

$$e^{x+iy} = e^x \cdot e^{iy} = e^x \cdot (\cos y + i \sin y) .$$

Notice that $e^{-iy} = \cos(-y) + i\sin(-y) = \cos y - i\sin y$ (we know that the sine function is odd and the cosine function even from their power series expansions). Then formula (9.29.1) tells us that

$$\cos y = \frac{e^{iy} + e^{-iy}}{2}$$

and

$$\sin y = \frac{e^{iy} - e^{-iy}}{2i}.$$

Now we may prove:

Proposition 9.31 *For every real x it holds that*

$$\sin^2 x + \cos^2 x = 1.$$

Proof: We see that

$$
\begin{aligned}
\sin^2 x + \cos^2 x &= \left(\frac{e^{ix} - e^{-ix}}{2i}\right)^2 + \left(\frac{e^{ix} + e^{-ix}}{2}\right)^2 \\
&= \frac{e^{2ix} - 2 + e^{-2ix}}{-4} + \frac{e^{2ix} + 2 + e^{-2ix}}{4} \\
&= 1.
\end{aligned}
$$

That completes the proof.   □

We list several other properties of the sine and cosine functions that may be proved by similar methods. The proofs are requested of you in the exercises.

Proposition 9.32 *The functions sine and cosine have the following properties:*

(a) $\sin(s + t) = \sin s \cos t + \cos s \sin t$;

(b) $\cos(s + t) = \cos s \cos t - \sin s \sin t$;

(c) $\cos(2s) = \cos^2 s - \sin^2 s$;

(d) $\sin(2s) = 2\sin s \cos s$;

(e) $\sin(-s) = -\sin s$;

(f) $\cos(-s) = \cos s$;

(g) $\sin'(s) = \cos s$;

(h) $\cos'(s) = -\sin s$.

One important task to be performed in a course on the foundations of analysis is to define the number π and establish its basic properties. In a course on Euclidean geometry, the constant π is defined to be the ratio of the circumference of a circle to its diameter. Such a definition is not useful for our purposes (however, it *is* consistent with the definition about to be given here).

Observe that $\cos 0$ is the real part of e^{i0} which is 1. Thus if we set

$$\alpha = \inf\{x > 0 : \cos x = 0\}$$

then $\alpha > 0$ and, by the continuity of the cosine function, $\cos \alpha = 0$. We define $\pi = 2\alpha$.

Applying Proposition 9.31 to the number α yields that $\sin \alpha = \pm 1$. Since α is the *first* zero of cosine on the right half line, the cosine function must be positive on $(0, \alpha)$. But cosine is the derivative of sine. Thus the sine function is *increasing* on $(0, \alpha)$. Since $\sin 0$ is the imaginary part of e^{i0} which is 0, we conclude that $\sin \alpha > 0$ hence that $\sin \alpha = +1$.

Now we may apply parts **(c)** and **(d)** of Proposition 9.25 with $s = \alpha$ to conclude that $\sin \pi = 0$ and $\cos \pi = -1$. A similar calculation with $s = \pi$ shows that $\sin 2\pi = 0$ and $\cos 2\pi = 1$. Next we may use parts **(a)** and **(b)** of Proposition 9.25 to calculate that $\sin(x + 2\pi) = \sin x$ and $\cos(x + 2\pi) = \cos x$ for all x. In other words, the sine and cosine functions are $2\pi-$periodic.

EXAMPLE 9.33 The business of calculating a decimal expansion for π would take us far afield. One approach would be to utilize the already-noted fact that the sine function is strictly increasing on the interval $[0, \pi/2]$ hence its inverse function

$$\text{Sin}^{-1} : [0, 1] \to [0, \pi/2]$$

is well defined. Then one can determine (see Chapter 6) that

$$\left(\text{Sin}^{-1}\right)'(x) = \frac{1}{\sqrt{1 - x^2}}.$$

By the Fundamental Theorem of Calculus,

$$\frac{\pi}{2} = \text{Sin}^{-1}(1) = \int_0^1 \frac{1}{\sqrt{1 - x^2}}\, dx.$$

By approximating the integral by its Riemann sums, one obtains an approximation to $\pi/2$ and hence to π itself. This approach will be explored in more detail in the exercises.

Let us for now observe that

$$\begin{aligned} \cos 2 &= 1 - \frac{2^2}{2!} + \frac{2^4}{4!} - \frac{2^6}{6!} + - \cdots \\ &= 1 - 2 + \frac{16}{24} - \frac{64}{720} + \cdots. \end{aligned}$$

Since the series defining $\cos 2$ is an alternating series with terms that strictly decrease to zero in magnitude, we may conclude (following reasoning from Chapter 4) that the last line is less than the sum of the first three terms:

$$\cos 2 < -1 + \frac{2}{3} < 0 \,.$$

It follows that $\alpha = \pi/2 < 2$ hence $\pi < 4$. A similar calculation of $\cos(3/2)$ would allow us to conclude that $\pi > 3$. $\qquad\square$

Exercises

1. Prove the equality $\left(\mathrm{Sin}^{-1}\right)' = 1/\sqrt{1 - x^2}$.

2. Prove that
$$\cos 2x = \cos^2 x - \sin^2 x$$
directly from the power series expansions.

3. Prove that
$$\sin 2x = 2 \sin x \cos x$$
directly from the power series expansions.

4. Use one of the methods described at the end of Section 3 to calculate π to two decimal places.

5. Prove that the trigonometric polynomials, that is to say, the functions of the form
$$p(x) = \sum_{j=-N}^{N} a_j e^{ijx} \,,$$
are dense in the continuous functions on $[0, 2\pi]$ in the uniform topology.

6. Find a formula for $\tan^4 x$ in terms of $\sin 2x$, $\sin 4x$, $\cos 2x$, and $\cos 4x$.

7. Prove Proposition 9.25(a), 9.25(b), 9.25(c).

8. Prove the general case of Proposition 9.29.

9. Derive a formula for $\cos 4x$ in terms of $\cos x$ and $\sin x$.

10. Provide the details of the assertion preceding Proposition 9.25 to the effect that if we define, for any real $\mathbb{R}$,
$$e^r = \sup\{q \in \mathbb{Q} : q < r\} \,,$$
then $e^x = \exp(x)$ for every real x.

11. Prove Proposition 9.32.

* **12.** Complete the following outline of a proof of Ivan Niven (see [NIV]) that π is irrational:

   (a) Define
   $$f(x) = \frac{x^n (1-x)^n}{n!},$$
   where n is a positive integer to be selected later. For each $0 < x < 1$ we have
   $$0 < f(x) < 1/n!. \tag{$*$}$$

   (b) For every positive integer j we have $f^{(j)}(0)$ is an integer.

   (c) $f(1-x) = f(x)$ hence $f^{(j)}(1)$ is an integer for every positive integer j.

   (d) Seeking a contradiction, assume that π is rational. Then π^2 is rational. Thus we may write $\pi^2 = a/b$, where a, b are positive integers and the fraction is in lowest terms.

   (e) Define
   $$F(x) = b^n \left(\pi^{2n} f(x) \right.$$
   $$- \pi^{2n-2} f^{(2)}(x) + \pi^{2n-4} f^{(4)}(x)$$
   $$\left. - \cdots + (-1)^n f^{(2n)}(x) \right).$$

   Then $F(0)$ and $F(1)$ are integers.

   (f) We have

   $$\tfrac{d}{dx} \left[F'(x) \sin(\pi x) \right.$$
   $$\left. - \pi F(x) \cos(\pi x) \right]$$
   $$= \pi^2 a^n f(x) \sin(\pi x).$$

   (g) We have
   $$\pi a^n \int_0^1 f(x) \sin(\pi x)\, dx$$
   $$= \left[\frac{F'(x) \sin x}{\pi} - F(x) \cos \pi x \right]_0^1$$
   $$= F(1) + F(0).$$

   (h) From this and $(*)$ we conclude that
   $$0 < \pi a^n \int_0^1 f(x) \sin(\pi x)\, dx$$
   $$< \frac{\pi a^n}{n!} < 1.$$
   When n is sufficiently large this contradicts the fact that $F(0)+F(1)$ is an integer.

9.4 Logarithms and Powers of Real Numbers

Definition 9.34 Since the exponential function $\exp(x) = e^x$ is positive and strictly increasing it is a one-to-one function from $\mathbb{R}$ to $(0, \infty)$. Thus it has a well-defined inverse function that we call the *natural logarithm*. We write this function as $\ln x$.

Proposition 9.35 *The natural logarithm function has the following properties:*

(a) $(\ln x)' = 1/x$;

(b) $\ln x$ *is strictly increasing;*

(c) $\ln(1) = 0$;

(d) $\ln e = 1$;

(e) *the graph of the natural logarithm function is asymptotic to the negative y axis;*

(f) $\ln(s \cdot t) = \ln s + \ln t$;

(g) $\ln(s/t) = \ln s - \ln t$.

Proof: These follow immediately from corresponding properties of the exponential function. For example, to verify part **(f)**, set $s = e^\sigma$ and $t = e^\tau$. Then

$$
\begin{aligned}
\ln(s \cdot t) &= \ln(e^\sigma \cdot e^\tau) \\
&= \ln(e^{\sigma + \tau}) \\
&= \sigma + \tau \\
&= \ln s + \ln t\,.
\end{aligned}
$$

The other parts of the proposition are proved similarly. □

Proposition 9.36 *If a and b are positive real numbers then*

$$
a^b = e^{b \cdot \ln a}\,.
$$

Proof: When b is an integer then the formula may be verified directly using Proposition 9.35, part **(f)**. For $b = m/n$ a rational number the formula follows by our usual trick of passing to nth roots. For arbitrary b we use a limiting argument as in our discussions of exponentials in Sections 2.3 and 9.3. □

EXAMPLE 9.37 We have discussed several different approaches to the exponentiation process. We proved the existence of nth roots, $n \in \mathbb{N}$, as an illustration of the completeness of the real numbers (by taking the supremum of a certain set). We treated rational exponents by composing the usual arithmetic process of taking mth powers with the process of taking nth roots. Then, in Sections 2.3 and 9.3, we passed to arbitrary powers by way of a limiting process.

Proposition 9.36 gives us a unified and direct way to treat all exponentials at once. This unified approach will prove (see the next proposition) to be particularly advantageous when we wish to perform calculus operations on exponential functions. □

Proposition 9.38 *Fix $a > 0$. The function $f(x) = a^x$ has the following properties:*

(a) $(a^x)' = a^x \cdot \ln a$;

(b) $f(0) = 1$;

(c) *if $0 < a < 1$ then f is decreasing and the graph of f is asymptotic to the positive x-axis;*

(d) *if $1 < a$ then f is increasing and the graph of f is asymptotic to the negative x-axis.*

Proof: These properties follow immediately from corresponding properties of the function exp. As an instance, to prove part **(a)**, we calculate that

$$(a^x)' = \left(e^{x \ln a}\right)' = e^{x \ln a} \cdot \ln a = a^x \cdot \ln a \,.$$

The other parts of the proposition are proved in a similar fashion. Details are left to the exercises. □

The logarithm function arises, among other places, in the context of probability and in the study of entropy. The reason is that the logarithm function is uniquely determined by the way that it interacts with the operation of multiplication:

Theorem 9.39 *Let $\phi(x)$ be a continuously differentiable function with domain the positive reals and which satisfies the identity*

$$\phi(s \cdot t) = \phi(s) + \phi(t) \tag{9.39.1}$$

for all positive s and t. Then there is a constant $C > 0$ such that

$$\phi(x) = C \cdot \ln x$$

for all x.

Proof: Differentiate the equation (9.39.1) with respect to s to obtain

$$t \cdot \phi'(s \cdot t) = \phi'(s) \,.$$

Now fix s and set $t = 1/s$ to conclude that

$$\phi'(1) \cdot \frac{1}{s} = \phi'(s) \,.$$

We take the constant C to be $\phi'(1)$ and apply Proposition 9.35(**a**) to conclude that $\phi(s) = C \cdot \ln s + D$ for some constant D. But ϕ cannot satisfy (9.39.1) unless $D = 0$, so the theorem is proved. $\qquad\qquad\square$

Observe that the *natural logarithm function* is then the unique continuously differentiable function that satisfies the condition (9.39.1) and whose derivative at 1 equals 1. That is the reason that the natural logarithm function (rather than the common logarithm, or logarithm to the base ten) is singled out as the focus of our considerations in this section.

Exercises

1. Calculate
$$\lim_{j \to \infty} \frac{j^{j/2}}{j!}.$$

2. At infinity, any nontrivial polynomial function dominates the natural logarithm function. Explain what this means, and prove it.

3. Give three distinct reasons why the natural logarithm function is not a polynomial.

4. Prove Proposition 9.38, parts (**b**), (**c**), (**d**), by following the hint provided.

5. Prove Proposition 9.35, except for part (**f**).

6. Prove that condition (9.39.1) implies that $\phi(1) = 0$. Assume that ϕ is differentiable at $x = 1$ but make no other hypothesis about the smoothness of ϕ. Prove that condition (9.39.1) then implies that ϕ is differentiable at every $x > 0$.

7. Show that the hypothesis of Theorem 9.39 may be replaced with $f \in \mathrm{Lip}_\alpha([0, 2\pi])$, some $\alpha > 0$.

8. Which function grows more quickly at infinity: $f(x) = x^k$ or $g(x) = |\ln x|^x$?

* 9. The *Lambert W function* is defined implicitly by the equation
$$z = W(z) \cdot e^{W(z)}.$$

It is a fact that any elementary transcendental function may be expressed (with an elementary formula) in terms of the W function. Prove that this is so for the exponential function and the sine function.

* 10. Prove Euler's formula relating the exponential to sine and cosine *not* by using power series, but rather by using differential equations.

Chapter 10

Applications of Analysis to Differential Equations

Differential equations are the heart and soul of analysis. Virtually any law of physics or engineering or biology or chemistry can be expressed as a differential equation—and frequently as a first-order equation (i.e., an equation involving only first derivatives). Much of mathematical analysis has been developed in order to find techniques for solving differential equations.

Most introductory books on differential equations ([COL], [KRA7], and [BIR] are three examples) devote themselves to elementary techniques for finding solutions to a very limited selection of equations. In the present book we take a different point of view. We instead explore certain central and broadly applicable principles which apply to virtually any differential equation. These principles, in particular, illustrate some of the key ideas of the book.

10.1 Picard's Existence and Uniqueness Theorem

10.1.1 The Form of a Differential Equation

A fairly general first-order differential equation will have the form

$$\frac{dy}{dx} = F(x, y) \,. \tag{10.1.1}$$

We say that the equation is "first order" because the highest derivative that appears is the first derivative.

In equation (10.1.1), F is a continuously differentiable function on some domain $(a, b) \times (c, d)$. We think of y as the dependent variable (that is, the function that we seek) and x as the independent variable. That is to say, $y = y(x)$. For technical reasons, we assume that the function F is bounded,

$$|F(x, y)| \le M \,, \tag{10.1.2}$$

and in addition that F satisfies a *Lipschitz condition*:

$$|F(x, s) - F(x, t)| \leq C \cdot |s - t| \,. \qquad (10.1.3)$$

[In many treatments it is standard to assume that F is bounded and $\partial F / \partial y$ is bounded. It is easy to see, using the Mean Value Theorem, that these two conditions imply (10.1.2), (10.1.3).]

EXAMPLE 10.1 Consider the equation

$$\frac{dy}{dx} = x^2 \sin y - y \ln x \,.$$

Then this equation fits the paradigm of equation (10.1.1) with $F(x, y) = x^2 \sin y - y \ln x$ provided that $1 \leq x \leq 2$ and $0 \leq y \leq 3$ (for instance). □

In fact the most standard, and physically appealing, setup for a first-order equation such as (10.1.1) is to adjoin to it an *initial condition*. For us this condition will have the form

$$y(x_0) = y_0 \,. \qquad (10.1.4)$$

Thus the problem we wish to solve is (10.1.1) and (10.1.4) together.

Picard's idea is to set up an iterative scheme for doing so. The most remarkable fact about Picard's technique is that it always works: As long as F is bounded and satisfies the Lipschitz condition, then the problem will possess one and only one solution.

10.1.2   Picard's Iteration Technique

While we will not actually give a complete proof that Picard's technique works, we will set it up and indicate the sequence of functions it produces that converges uniformly to the solution of our problem.

Picard's approach is inspired by the fact that the differential equation (10.1.1) and initial condition (10.1.4), taken together, are equivalent to the single integral equation

$$y(x) = y_0 + \int_{x_0}^{x} F(t, y(t)) \, dt \,. \qquad (10.1.5)$$

We invite the reader to differentiate both sides of this equation, using the Fundamental Theorem of Calculus, to derive the original differential equation (10.1.1). Of course the initial condition (10.1.4) is built into (10.1.5). This integral equation inspires the iteration scheme that we now describe.

We assume that $x_0 \in (a, b)$ and that $y_0 \in (c, d)$. We set

$$y_1(x) = y_0 + \int_{x_0}^{x} F(t, y_0) \, dt \,.$$

For x near to x_0, this definition makes sense.

Next we define

$$y_2(x) = y_0 + \int_{x_0}^{x} F(t, y_1(t)) \, dt$$

and, more generally,

$$y_{j+1}(x) = y_0 + \int_{x_0}^{x} F(t, y_j(t)) \, dt \qquad (10.1.6)$$

for $j = 2, 3, \ldots$.

It turns out that the sequence of functions $\{y_1, y_2, \ldots\}$ will converge uniformly on an interval of the form $[x_0 - h, x_0 + h] \subseteq (a, b)$ to a solution of (10.1.1) that satisfies (10.1.4).

10.1.3   Some Illustrative Examples

Picard's iteration method is best apprehended by way of some examples that show how the iterates arise and how they converge to a solution. We now proceed to develop such illustrations.

EXAMPLE 10.2 Consider the initial value problem

$$y' = 2y \, , \qquad y(0) = 1 \, .$$

Of course this could easily be solved by the method of first order linear equations, or by separation of variables (see [KRA7] for a description of these methods). Our purpose here is to illustrate how the Picard method works.

First notice that the stated initial value problem is equivalent to the integral equation

$$y(x) = 1 + \int_{0}^{x} 2y(t) \, dt \, .$$

Following the paradigm (10.1.6), we thus find that

$$y_{j+1}(x) = 1 + \int_{0}^{x} 2y_j(x) \, dx \, .$$

Using $x_0 = 0$, $y_0 = 1$, we then find that

$$
\begin{aligned}
y_1(x) &= 1 + \int_{0}^{x} 2 \, dt = 1 + 2x \, , \\
y_2(x) &= 1 + \int_{0}^{x} 2(1 + 2t) \, dt = 1 + 2x + 2x^2 \, , \\
y_3(x) &= 1 + \int_{0}^{x} 2(1 + 2t + 2t^2) \, dt = 1 + 2x + 2x^2 + \frac{4x^3}{3} \, .
\end{aligned}
$$

In general, we find that

$$y_j(x) = 1 + \frac{2x}{1!} + \frac{(2x)^2}{2!} + \frac{(2x)^3}{3!} + \cdots + \frac{(2x)^j}{j!} = \sum_{\ell=0}^{j} \frac{(2x)^\ell}{\ell!} \, .$$

It is plain that these are the partial sums for the power series expansion of $y = e^{2x}$. We conclude that the solution of our initial value problem is $y = e^{2x}$. You are encouraged to check that $y = e^{2x}$ does indeed solve the differential equation and initial condition stated at the beginning of the example.   □

EXAMPLE 10.3 Let us use Picard's method to solve the initial value problem

$$y' = 2x - y , \qquad y(0) = 1 .$$

The equivalent integral equation is

$$y(x) = 1 + \int_0^x [2t - y(t)] \, dt$$

and (10.1.6) tells us that

$$y_{j+1}(x) = 1 + \int_0^x [2t - y_j(t)] \, dt .$$

Taking $x_0 = 0$, $y_0 = 1$, we then find that

$$
\begin{aligned}
y_1(x) &= 1 + \int_0^x (2t - 1) \, dt = 1 + x^2 - x , \\
y_2(x) &= 1 + \int_0^x \left(2t - [1 + t^2 - t] \right) dt \\
&= 1 + \frac{3x^2}{2} - x - \frac{x^3}{3} , \\
y_3(x) &= 1 + \int_0^x \left(2t - [1 + 3t^2/2 - t - t^3/3] \right) dt \\
&= 1 + \frac{3x^2}{2} - x - \frac{x^3}{2} + \frac{x^4}{4 \cdot 3} , \\
y_4(x) &= 1 + \int_0^x \left(2t - [1 + 3t^2/2 - t - t^3/2 + t^4/4 \cdot 3] \right) dt \\
&= 1 + \frac{3x^2}{2} - x - \frac{x^3}{2} + \frac{x^4}{4 \cdot 2} - \frac{x^5}{5 \cdot 4 \cdot 3} .
\end{aligned}
$$

In general, we find that

$$
\begin{aligned}
y_j(x) &= 1 - x + \frac{3x^2}{2!} - \frac{3x^3}{3!} + \frac{3x^4}{4!} - + \cdots \\
&\quad + (-1)^j \frac{3x^j}{j!} + (-1)^{j+1} \frac{2x^{j+1}}{(j+1)!} \\
&= [2x - 2] + 3 \cdot \sum_{\ell=0}^{j} \frac{(-x)^j}{j!} + (-1)^{j+1} \frac{2x^{j+1}}{(j+1)!} .
\end{aligned}
$$

Notice that the $2x - 2$ terms cancel with the first two terms of the infinite sum to give $1 - x$.

Of course the last term tends to 0 as $j \to +\infty$. Thus we see that the iterates $y_j(x)$ converge to the solution $y(x) = [2x-2]+3e^{-x}$ for the initial value problem. Check that this function does indeed satisfy the given differential equation and initial condition. $\square$

10.1.4   Estimation of the Picard Iterates

To get an idea of why the assertion at the end of Subsection 10.1.2—that the functions y_j converge uniformly—is true, let us do some elementary estimations. Choose $h > 0$ so small that $h \cdot C < 1$, where C is the constant from the Lipschitz condition (10.1.3). We will assume in the following calculations that $|x-x_0| < h$.

Now we proceed with the iteration. Let y_0 be the initial value at x_0 as usual. Then

$$\begin{aligned}
|y_0 - y_1(t)| &= \left| \int_{x_0}^{x} F(t, y_0)\, dt \right| \\
&\leq \int_{x_0}^{x} |F(t, y_0)|\, dt \\
&\leq M \cdot |x - x_0| \\
&\leq M \cdot h\,.
\end{aligned}$$

We have of course used the boundedness condition (10.1.2).

Next we have

$$\begin{aligned}
|y_1(x) - y_2(x)| &= \left| \int_{x_0}^{x} F(t, y_0(t))\, dt - \int_{x_0}^{x} F(t, y_1(t))\, dt \right| \\
&\leq \int_{x_0}^{x} |F(t, y_0(t)) - F(t, y_1(t))|\, dt \\
&\leq \int_{x_0}^{x} C \cdot |y_0(t) - y_1(t)|\, dt \\
&\leq C \cdot M \cdot h \cdot h \\
&= M \cdot C \cdot h^2\,.
\end{aligned}$$

One can continue this procedure to find that

$$|y_2(x) - y_3(x)| \leq M \cdot C^2 \cdot h^3 = M \cdot h \cdot (Ch)^2\,.$$

and, more generally,

$$|y_j(x) - y_{j+1}(x)| \leq M \cdot C^j \cdot h^{j+1} = M \cdot h \cdot (Ch)^j\,.$$

Now, if $0 < K < L$ are integers, then

$$\begin{aligned}
|y_K(x) - y_L(x)| &\leq |y_K(x) - y_{K+1}(x)| + |y_{K+1}(x) - y_{K+2}(x)| \\
&\quad + \cdots + |y_{L-1}(x) - y_L(x)| \\
&\leq M \cdot h \cdot \left([Ch]^K + [Ch]^{K+1} + \cdots [Ch]^{L-1} \right)\,.
\end{aligned}$$

Since $|Ch| < 1$ by design, the geometric series $\sum_j [Ch]^j$ converges. As a result, the expression on the right of our last display is as small as we please, for K and L large, just by the Cauchy criterion for convergent series. It follows that the sequence $\{y_j\}$ of approximate solutions converges uniformly to a function $y = y(x)$. In particular, y is continuous.

Furthermore, we know that

$$y_{j+1}(x) = y_0 + \int_{x_0}^x F(t, y_j(t))\, dt \,.$$

Letting $j \to \infty$, and invoking the uniform convergence of the y_j, we may pass to the limit and find that

$$y(x) = y_0 + \int_{x_0}^x F(t, y(x))\, dt \,.$$

This says that y satisfies the integral equation that is equivalent to our original initial value problem. This equation also shows that y is continuously differentiable. Thus y is the function that we seek.

It can be shown that this y is in fact the *unique* solution to our initial value problem. We shall not provide the details of the proof of that assertion.

In case F is not Lipschitz—say that F is only continuous—then it is still possible to show that a solution y exists. But it will no longer be unique.

Exercises

1. Use the method of Picard iteration to solve the initial value problem $y' = y + x$, $y(0) = 1$.

2. Verify that the function $y = 1/\sqrt{2(x+1)}$ is a solution of the differential equation

   $$y' + y^3 = 0 \,. \tag{$*$}$$

   Can you use separation of variables to find the general solution? This means to write the equation as

   $$\frac{dy}{dx} = -y^3$$

   and then do some algebra to have all x terms on one side of the equation and all y terms on the other side of the equation. Then integrate.

   [**Hint:** It is $y = 1/\sqrt{2(x+c)}$.] Now find the solution to the initial value problem $(*)$ with initial condition $y(1) = 4$.

3. Check that the function

   $$y = \sqrt{\frac{2}{3}\ln(1 + x^2) + C}$$

solves the differential equation

$$\frac{dy}{dx} = \frac{2x}{3y + 3yx^2}.$$

Find the particular solution that satisfies the initial condition $y(0) = 2$.

4. In the method of Picard, suppose that the function F is given by a power series. Formulate a version of the Picard iteration technique in the language of power series.

5. Explain why the initial value problem

$$\begin{aligned} y' &= e^y \\ y(0) &= 1 \end{aligned}$$

has a solution in a neighborhood of the origin.

6. For each differential equation, sketch the family of solutions on a set of axes. This means, since each equation is *not* equipped with an initial condition, that the solution to each equation will have an unspecified constant in it.

(a) $y' - xy = 0$
(b) $y' + y = e^x$
(c) $y' = x$
(d) $y' = 1 - y$

* **7.** Formulate a version of the Picard theorem for vector-valued functions. Indicate how its proof differs, if at all, from the proof for scalar-valued functions. Now explain how one can use this vector-valued version of Picard to obtain an existence and uniqueness theorem for kth-order ordinary differential equations.

* **8.** Does the Picard theorem apply to the initial value problem

$$e^{dy/dx} + \frac{dy}{dx} = x^2, \quad y(1) = 2?$$

Why or why not? [**Hint:** Think in terms of the Implicit Function Theorem.]

* **9.** A *vector field* is a function

$$F(x, y) = \langle \alpha(x, y), \beta(x, y) \rangle$$

that assigns to each point in the plane $\mathbb{R}^2$ a vector. We call a curve $\gamma : (a, b) \to \mathbb{R}^2$ (here $\gamma(t) = (\gamma_1(t), \gamma_2(t))$) an *integral curve* of the vector field if

$$\gamma'(t) = F(\gamma(t))$$

for each t. Thus γ "flows along" the vector field, and the tangent to the curve at each point is given by the value of the vector field at that point.

Put suitable conditions on F that will guarantee that if $P \in \mathbb{R}^2$ then there will be an integral curve for F through the point P. [**Hint:** Of course use the Picard theorem to obtain your result. What is the correct initial value problem?]

* **10.** Give an example which illustrates that the integral curve that you found in Exercise **9** will only, in general, be defined in a small neighborhood of P. [**Hint:** Think of a vector field that "dies out."]

* **11.** Refer to Exercises **9** and **10**. Find integral curves for each of the following vector fields:

    (a) $F(x, y) = \langle -y, x \rangle$
    (b) $F(x, y) = \langle x + 1, y - 2 \rangle$
    (c) $F(x, y) = \langle 2xy, x^2 \rangle$
    (d) $F(x, y) = \langle -x, 2y \rangle$

10.2   Power Series Methods

One of the techniques of broadest applicability in the subject of differential equations is that of power series, or real analytic functions. The philosophy is to *guess* that a given problem has a solution that may be represented by a power series, and then to endeavor to solve for the coefficients of that series. Along the way, one uses (at least tacitly) fundamental properties of these series—that they may be differentiated and integrated term by term, for instance. And that their intervals of convergence are preserved under standard arithmetic operations.

EXAMPLE 10.4 Let p be an arbitrary real constant. Let us use a differential equation to derive the power series expansion for the function

$$y = (1 + x)^p .$$

Of course the given y is a solution of the initial value problem

$$(1 + x) \cdot y' = py , \quad y(0) = 1 .$$

We assume that the equation has a power series solution

$$y = \sum_{j=0}^{\infty} a_j x^j = a_0 + a_1 x + a_2 x^2 + \cdots$$

with positive radius of convergence R. Then

$$y' = \sum_{j=1}^{\infty} j \cdot a_j x^{j-1} = a_1 + 2a_2 x + 3a_3 x^2 + \cdots ;$$

$$xy' = \sum_{j=1}^{\infty} j \cdot a_j x^j = a_1 x + 2a_2 x^2 + 3a_3 x^3 + \cdots ;$$

$$py = \sum_{j=0}^{\infty} pa_j x^j = pa_0 + pa_1 x + pa_2 x^2 + \cdots .$$

By the differential equation, we see that the sum of the first two of these series equals the third. Thus

$$\sum_{j=1}^{\infty} ja_j x^{j-1} + \sum_{j=1}^{\infty} ja_j x^j = \sum_{j=0}^{\infty} pa_j x^j .$$

We immediately see two interesting anomalies: the powers of x on the left-hand side do not match up, so the two series cannot be immediately added. Also the summations do not all begin in the same place. We address these two concerns as follows.

First, we can change the index of summation in the first sum on the left to obtain

$$\sum_{j=0}^{\infty} (j+1)a_{j+1} x^j + \sum_{j=1}^{\infty} ja_j x^j = \sum_{j=0}^{\infty} pa_j x^j .$$

Write out the first few terms of the new sum, and the original sum, to see that they are just the same.

Now every one of our series has x^j in it, but they begin at different places. So we break off the extra terms as follows:

$$\sum_{j=1}^{\infty} (j+1)a_{j+1} x^j + \sum_{j=1}^{\infty} ja_j x^j - \sum_{j=1}^{\infty} pa_j x^j = -a_1 x^0 + pa_0 x^0 . \qquad (10.5.1)$$

Notice that all we have done is to break off the zeroeth terms of the first and third series, and put them on the right.

The three series on the left-hand side of (10.5.1) are begging to be put together: they have the same form, they all involve powers of x, and they all begin at the same index. Let us do so:

$$\sum_{j=1}^{\infty} [(j+1)a_{j+1} + ja_j - pa_j] x^j = -a_1 + pa_0 .$$

Now the powers of x that appear on the left are 1, 2, ..., and there are none of these on the right. We conclude that each of the coefficients on the left is zero; by the same reasoning, the coefficient $(-a_1 + pa_0)$ on the right (i.e., the constant term) equals zero. So we have the equations[1]

$$-a_1 + pa_0 = 0$$
$$(j+1)a_{j+1} + (j-p)a_j = 0 \quad \text{for } j \geq 1 .$$

---

[1]A set of equations like this is called a *recursion*. It expresses a_js with later indices in terms of a_js with earlier indices.

Our initial condition tells us that $a_0 = 1$. Then our first equation implies that $a_1 = p$. The next equation, with $j = 1$, says that

$$2a_2 + (1 - p)a_1 = 0 \, .$$

Hence $a_2 = (p - 1)a_1/2 = (p - 1)p/2$. Continuing, we take $j = 2$ in the second equation to get

$$3a_3 + (2 - p)a_2 = 0$$

so $a_3 = (p - 2)a_2/3 = (p - 2)(p - 1)p/(3 \cdot 2)$.

We may continue in this manner to obtain that

$$a_j = \frac{p(p - 1)(p - 2) \cdots (p - j + 1)}{j!} \quad \text{for } j \geq 1 \, .$$

Thus the power series expansion for our solution y is

$$y \quad = \quad 1 + px + \frac{p(p - 1)}{2!}x^2 + \frac{p(p - 1)(p - 2)}{3!}x^3 + \cdots$$
$$+ \frac{p(p - 1)(p - 2) \cdots (p - j + 1)}{j!}x^j + \cdots \, .$$

Since we knew in advance that the solution of our initial value problem was

$$y = (1 + x)^p \, ,$$

we find that we have derived Isaac Newton's general binomial theorem (or binomial series):

$$(1 + x)^p \quad = \quad 1 + px + \frac{p(p - 1)}{2!}x^2 + \frac{p(p - 1)(p - 2)}{3!}x^3 + \cdots$$
$$+ \frac{p(p - 1)(p - 2) \cdots (p - j + 1)}{j!}x^j + \cdots \, .$$

□

EXAMPLE 10.5 Let us consider the differential equation

$$y' = y \, .$$

Of course we know from elementary considerations that the solution to this equation is $y = C \cdot e^x$, but let us pretend that we do not know this. Our goal is to instead use power series to *discover* the solution. We proceed by *guessing* that the equation has a solution given by a power series, and we proceed to solve for the coefficients of that power series.

So our guess is a solution of the form

$$y = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \cdots \, .$$

Then

$$y' = a_1 + 2a_2 x + 3a_3 x^2 + \cdots \, ,$$

and we may substitute these two expressions into the differential equation. Thus

$$a_1 + 2a_2 x + 3a_3 x^2 + \cdots = a_0 + a_1 x + a_2 x^2 + \cdots .$$

Now the powers of x must match up (i.e., the coefficients must be equal). We conclude that

$$
\begin{aligned}
a_1 &= a_0 \\
2a_2 &= a_1 \\
3a_3 &= a_2
\end{aligned}
$$

and so forth. Let us take a_0 to be an unknown constant C. Then we see that

$$
\begin{aligned}
a_1 &= C; \\
a_2 &= \frac{C}{2}; \\
a_3 &= \frac{C}{3 \cdot 2};
\end{aligned}
$$

etc.

In general,

$$a_j = \frac{C}{j!} .$$

In summary, our power series solution of the original differential equation is

$$y = \sum_{j=0}^{\infty} \frac{C}{j!} x^j = C \cdot \sum_{j=0}^{\infty} \frac{x^j}{j!} = C \cdot e^x .$$

Thus we have a new way, using power series, of discovering the general solution of the differential equation $y' = y$. $\quad\square$

EXAMPLE 10.6 Let us use the method of power series to solve the differential equation

$$(1 - x^2)y'' - 2xy' + p(p+1)y = 0 . \tag{10.7.1}$$

Here p is an arbitrary real constant. This is called *Legendre's equation*.

We therefore guess a solution of the form

$$y = \sum_{j=0}^{\infty} a_j x^j = a_0 + a_1 x + a_2 x^2 + \cdots$$

and calculate

$$y' = \sum_{j=1}^{\infty} j a_j x^{j-1} = a_1 + 2a_2 x + 3a_3 x^2 + \cdots$$

and

$$y'' = \sum_{j=2}^{\infty} j(j-1)a_j x^{j-2} = 2a_2 + 3 \cdot 2 \cdot a_3 x + \cdots .$$

It is most convenient to treat the differential equation in the form (10.7.1). We calculate

$$-x^2 y'' = -\sum_{j=2}^{\infty} j(j-1)a_j x^j$$

and

$$-2xy' = -\sum_{j=1}^{\infty} 2j a_j x^j .$$

Substituting into the differential equation now yields

$$\sum_{j=2}^{\infty} j(j-1)a_j x^{j-2} - \sum_{j=2}^{\infty} j(j-1)a_j x^j - \sum_{j=1}^{\infty} 2j a_j x^j + p(p+1)\sum_{j=0}^{\infty} a_j x^j = 0 .$$

We adjust the index of summation in the first sum so that it contains x^j rather than x^{j-2} and we break off spare terms and collect them on the right. We also break off terms from the third and fourth power series and move them to the right. The result is

$$\sum_{j=2}^{\infty} (j+2)(j+1)a_{j+2} x^j - \sum_{j=2}^{\infty} j(j-1)a_j x^j$$

$$-\sum_{j=2}^{\infty} 2j a_j x^j + p(p+1)\sum_{j=2}^{\infty} a_j x^j$$

$$= -2a_2 - 6a_3 x + 2a_1 x - p(p+1)a_0 - p(p+1)a_1 x .$$

In other words,

$$\sum_{j=2}^{\infty} \left[(j+2)(j+1)a_{j+2} - j(j-1)a_j - 2j a_j + p(p+1)a_j \right] x^j$$

$$= -2a_2 - 6a_3 x + 2a_1 x - p(p+1)a_0 - p(p+1)a_1 x .$$

As a result,

$$\left[(j+2)(j+1)a_{j+2} - j(j-1)a_j - 2j a_j + p(p+1)a_j \right] = 0 \qquad \text{for } j = 2, 3, \ldots$$

together with

$$-2a_2 - p(p+1)a_0 = 0$$

and

$$-6a_3 + 2a_1 - p(p+1)a_1 = 0 .$$

We have arrived at the recursion

$$a_2 = -\frac{p(p+1)}{1 \cdot 2} \cdot a_0 \,,$$

$$a_3 = -\frac{(p-1)(p+2)}{2 \cdot 3} \cdot a_1 \,,$$

$$a_{j+2} = -\frac{(p-j)(p+j+1)}{(j+2)(j+1)} \cdot a_j \qquad \text{for } j = 2, 3, \dots. \qquad (10.7.2)$$

We recognize a familiar pattern: The coefficients a_0 and a_1 are unspecified, so we set $a_0 = A$ and $a_1 = B$. Then we may proceed to solve for the rest of the coefficients. Now

$$a_2 = -\frac{p(p+1)}{2} \cdot A \,,$$

$$a_3 = -\frac{(p-1)(p+2)}{2 \cdot 3} \cdot B \,,$$

$$a_4 = -\frac{(p-2)(p+3)}{3 \cdot 4} a_2 = \frac{p(p-2)(p+1)(p+3)}{4!} \cdot A \,,$$

$$
\begin{aligned}
a_5 &= -\frac{(p-3)(p+4)}{4 \cdot 5} a_3 \\
&= \frac{(p-1)(p-3)(p+2)(p+4)}{5!} \cdot B \,,
\end{aligned}
$$

$$
\begin{aligned}
a_6 &= -\frac{(p-4)(p+5)}{5 \cdot 6} a_4 \\
&= -\frac{p(p-2)(p-4)(p+1)(p+3)(p+5)}{6!} \cdot A \,,
\end{aligned}
$$

$$
\begin{aligned}
a_7 &= -\frac{(p-5)(p+6)}{6 \cdot 7} a_5 \\
&= -\frac{(p-1)(p-3)(p-5)(p+2)(p+4)(p+6)}{7!} \cdot B \,,
\end{aligned}
$$

and so forth. Putting these coefficient values into our supposed power series solution we find that the general solution of our differential equation is

$$
\begin{aligned}
y \;=\; & A\Bigl[1 - \frac{p(p+1)}{2!}x^2 + \frac{p(p-2)(p+1)(p+3)}{4!}x^4 \\
& - \frac{p(p-2)(p-4)(p+1)(p+3)(p+5)}{6!}x^6 + - \cdots \Bigr] \\
& + B\Bigl[x - \frac{(p-1)(p+2)}{3!}x^3 + \frac{(p-1)(p-3)(p+2)(p+4)}{5!}x^5 \\
& - \frac{(p-1)(p-3)(p-5)(p+2)(p+4)(p+6)}{7!}x^7 + - \cdots \Bigr].
\end{aligned}
$$

We assure the reader that, when p is not an integer, then these are *not* familiar elementary transcendental functions. They are what we call *Legendre functions*. In the special circumstance that p is a positive even integer, the first function (that which is multiplied by A) terminates as a polynomial. In the special circumstance that p is a positive odd integer, the second function (that which is multiplied by B) terminates as a polynomial. These are called *Legendre polynomials*, and they play an important role in mathematical physics, representation theory, and interpolation theory. □

Some differential equations have singularities. In the present context, this means that the higher order terms have coefficients that vanish to high degree. As a result, one must make a slightly more general guess as to the solution of the equation. This more general guess allows for a corresponding singularity to be built into the solution. Rather than develop the full theory of these Frobenius series, we merely give one example.

EXAMPLE 10.7 We use the method of Frobenius series to solve the differential equation

$$2x^2 y'' + x(2x + 1)y' - y = 0 \tag{10.8.1}$$

about the regular singular point 0.

We guess a solution of the form

$$y = x^m \cdot \sum_{j=0}^{\infty} a_j x^j = \sum_{j=0}^{\infty} a_j x^{m+j}$$

and therefore calculate that

$$y' = \sum_{j=0}^{\infty} (m+j)a_j x^{m+j-1}$$

and

$$y'' = \sum_{j=0}^{\infty} (m+j)(m+j-1)a_j x^{m+j-2} \, .$$

Substituting these calculations into the differential equation yields

$$2\sum_{j=0}^{\infty} (m+j)(m+j-1)a_j x^{m+j}$$

$$+2\sum_{j=0}^{\infty} (m+j)a_j x^{m+j+1}$$

$$+\sum_{j=0}^{\infty} (m+j)a_j x^{m+j} - \sum_{j=0}^{\infty} a_j x^{m+j}$$

$$= 0 \, .$$

We make the usual adjustments in the indices so that all powers of x are x^{m+j}, and break off the odd terms to put on the right-hand side of the equation. We obtain

$$2\sum_{j=1}^{\infty}(m+j)(m+j-1)a_j x^{m+j}$$

$$+2\sum_{j=1}^{\infty}(m+j-1)a_{j-1}x^{m+j}$$

$$+\sum_{j=1}^{\infty}(m+j)a_j x^{m+j} - \sum_{j=1}^{\infty}a_j x^{m+j}$$

$$= -2m(m-1)a_0 x^m - ma_0 x^m + a_0 x^m \,.$$

The result is

$$\left[2(m+j)(m+j-1)a_j + 2(m+j-1)a_{j-1}\right.$$

$$\left. +(m+j)a_j - a_j\right] = 0$$

$$\text{for } j = 1, 2, 3, \ldots \tag{10.8.2}$$

together with

$$[-2m(m-1) - m + 1]a_0 = 0\,.$$

It is clearly not to our advantage to let $a_0 = 0$. Thus

$$-2m(m-1) - m + 1 = 0\,.$$

This is the *indicial equation*.

The roots of this quadratic equation are $m = -1/2, 1$. We put each of these values into (10.8.2) and solve the resulting recursion.

Now (10.8.2) says that

$$(2m^2 + 2j^2 + 4mj - j - m - 1)a_j = (-2m - 2j + 2)a_{j-1}\,.$$

For $m = -1/2$ this is

$$a_j = \frac{3 - 2j}{-3j + 2j^2}a_{j-1}$$

so

$$a_1 = -a_0\,, \quad a_2 = -\frac{1}{2}a_1 = \frac{1}{2}a_0\,, \text{ etc.}$$

For $m = 1$ we have

$$a_j = \frac{-2j}{3j + 2j^2}a_{j-1}$$

so

$$a_1 = -\frac{2}{5}a_0\,, \quad a_2 = -\frac{4}{14}a_1 = \frac{4}{35}a_0\,.$$

Thus we have found the linearly independent solutions

$$a_0 x^{-1/2} \cdot \left(1 - x + \frac{1}{2}x^2 - + \cdots\right)$$

and

$$a_0 x \cdot \left(1 - \frac{2}{5}x + \frac{4}{35}x^2 - + \cdots\right).$$

The general solution of our differential equation is then

$$y = A x^{-1/2} \cdot \left(1 - x + \frac{1}{2}x^2 - + \cdots\right) + B x \cdot \left(1 - \frac{2}{5}x + \frac{4}{35}x^2 - + \cdots\right).$$

<div align="right">□</div>

Exercises

1. Explain why the method of power series would not work very well to solve the differential equation

$$y' - |x|y = \sin x.$$

Note here that the coefficient of y is $|x|$, and $|x|$ is not a differentiable function.

2. Solve the initial value problem

$$y'' - xy = x^2 \quad , \quad y(0) = 2, \ y'(0) = 1$$

by the method of power series.

3. Solve the initial value problem

$$y' - xy = x \quad , \quad y(0) = 2$$

by the method of power series.

4. Solve the differential equation

$$y''' - xy' = x$$

by the method of power series. Since there are no initial conditions, you should obtain a general solution with three free parameters.

5. Solve the initial value problem

$$y' - y = x \quad , \quad y(0) = 1$$

both by Picard's method and by the method of power series. Verify that you get the same solution by both means.

6. When you solve a differential equation by the method of power series, you cannot in general expect the power series to converge on the entire real line. As an example, solve the differential equation

$$y' - \frac{1}{1-x}y = 0$$

by the method of power series. What is the radius of convergence of the power series? Can you suggest why that is so?

7. Consider the differential equation

$$y'' - y = x^2 \,.$$

The function x^2 is even. If the function y is even, then y'' will be even also. Thus it makes sense to suppose that there is a power series solution with only even powers of x. Find it.

8. Consider the differential equation

$$y'' + y = x^3 \,.$$

The function x^3 is odd. If the function y is odd, then y'' will also be odd. Thus it makes sense to suppose that there is a power series solution with only odd powers of x. Find it.

9. Find all solutions of the differential equation

$$y' = xy \,.$$

10. Find all solutions of the differential equation

$$y' = \frac{y}{x} \,.$$

11. Use power series methods to solve the differential equation

$$y'' + 4y = 0 \,.$$

12. Solve the differential equation

$$y' = y^2 \,.$$

* 13. What are sufficient conditions on the function F so that the differential equation

$$y' = F(x, y)$$

has the property that its solution y is continuously differentiable?

* 14. Find a solution of the partial differential equation

$$\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} \right) u(x, y) = x + y$$

using the method of power series in two variables.

Chapter 11

Introduction to Harmonic Analysis

11.1 The Idea of Harmonic Analysis

Fourier analysis first arose historically in the context of the study of a certain partial differential equation (we shall describe this equation in detail in the discussion below) of mathematical physics. The equation could be solved explicitly when the input (i.e., the right-hand side of the equation) was a function of the form $\sin jx$ or $\cos jx$ for j an integer. The question arose whether an *arbitrary* input could be realized as the superposition of sine functions and cosine functions.

In the late eighteenth century, debate raged over this question. It was fueled by the fact that there was no solid understanding of just what constituted a function. The important treatise [FOU] of Joseph Fourier gave a somewhat dreamy but nevertheless precise method for expanding virtually any function as a series in sines and cosines. It took almost a century, and the concerted efforts of Dirichlet, Cauchy, Riemann, Weierstrass, and many other important analysts, to put the so-called theory of "Fourier series" on a rigorous footing.

We now know, and can prove exactly, that if f is a continuously differentiable function on the interval $[0, 2\pi]$ then the coefficients

$$c_n = \frac{1}{2\pi} \int_0^{2\pi} f(t)e^{-int}\, dt$$

give rise to a series expansion

$$f(t) = \sum_{n=0}^{\infty} c_n e^{int}$$

that is valid (i.e., convergent) at every point, and converges back to f. [Notice that the convenient notation e^{ijt} given to us by Euler's formula carries

information both about the sine and the cosine.] This expansion validates the vague but aggressive ruminations in [FOU] and lays the foundations for a powerful and deep method of analysis that today has wide applicability in physics, engineering, differential equations, and harmonic analysis.[1]

In the present chapter we shall explore the foundations of Fourier series and also learn some of their applications. All of our discussions will of course be rigorous and precise. They will take advantage of all the tools of analysis that we have developed thus far in the present book.

Exercises

1. The function $f(\theta) = \cos^4 \theta$ is a nice smooth function, so will have a Fourier series expansion. That is, it will have an expansion as a sum of functions $\cos j\theta$ and $\sin j\theta$ with real coefficients. Determine what that expansion is.

* 2. Explain why the only continuous multiplicative homomorphisms from the circle group $\mathbb{T}$, which is just the set of all $e^{i\theta}$ in the plane, into $\mathbb{C} \setminus \{0\}$ are given by

$$e^{i\theta} \longmapsto e^{ik\theta}$$

for some integer k. Here a homomorphism φ in this context is a function φ that satisfies $\varphi(a \cdot b) = \varphi(a) \cdot \varphi(b)$.

* 3. Answer Exercise **2** with the circle group replaced by the real line.

4. Classical harmonic analysis is done on a space with a group action—such as the circle group, or the line, or N-dimensional Euclidean space. Explain what this assertion means, and supply some detail.

* 5. It can be proved, using elementary Fourier series (see Section 11.2), that

$$\sum_{j=1}^{\infty} \frac{1}{j^2} = \frac{\pi^2}{6}.$$

This fact was established by Leonhard Euler in 1735. It is a matter of great interest to find similar formulas for

$$\sum_{j=1}^{\infty} \frac{1}{j^k}$$

when $k = 3, 4, \ldots$. Apery has shown that, when $k = 3$, then the sum is irrational. This set of ideas has to do with the Riemann zeta function and the distribution of primes. Do some experiments on your computer to determine what this might mean.

---

[1]Notice that the result enunciated here is a decisive improvement over what we know about Taylor series. We have asserted that a function that is only continuously differentiable has a Fourier series that converges at every point. But even an infinitely differentiable function can have Taylor series that converges at no point.

6. Refer to Exercise 5. Use your symbol manipulation software to calculate the partial sums S_{100}, S_{1000}, and S_{10000} for the series

$$\sum_{j=1}^{\infty} \frac{1}{j^2} .$$

Compare your answers with the value of $\pi^2/6$.

11.2 The Elements of Fourier Series

In this section it will be convenient for us to work on the interval $[0, 2\pi]$. We will perform arithmetic operations on this interval *modulo* 2π: for example, $3\pi/2 + 3\pi/2$ is understood to equal π because we subtract from the answer the largest multiple of 2π that it exceeds. When we refer to a function f being continuous on $[0, 2\pi]$, we require that it be right continuous at 0, left continuous at 2π, and that $f(0) = f(2\pi)$. Similarly for continuous differentiability and so forth.

If f is a (either real- or complex-valued) Riemann integrable function on this interval and if $n \in \mathbb{Z}$, then we define

$$\widehat{f}(n) = \frac{1}{2\pi} \int_0^{2\pi} f(t)e^{-int} \, dt .$$

We call $\widehat{f}(n)$ the nth *Fourier coefficient* of f. The formal expression

$$Sf(x) \sim \sum_{n=-\infty}^{\infty} \widehat{f}(n)e^{inx}$$

is called the *Fourier series* of the function f. Notice that we are *not* claiming that Sf converges, nor that it converges to f. Right now it is just a formal expression.

In circumstances where the Fourier series converges to the function f, some of which we shall discuss below, the series provides a decomposition of f into simple component functions. This type of analysis is of importance in the theory of differential equations, in signal and image processing, and in scattering theory. There is a rich theory of Fourier series which is of interest in its own right.

It is important that we say right away how we sum Fourier series. Define the Nth *partial sum* of the Fourier series of f to be

$$S_N f(x) = \sum_{j=-N}^{N} \widehat{f}(j)e^{ijx} .$$

We say that the Fourier series Sf *converges* to f at x if $S_N f(x) \to f(x)$.

Observe that, in case f has the special form

$$f(x) = \sum_{j=-N}^{N} a_j e^{ijt} , \tag{11.2.1}$$

then we may calculate that

$$\frac{1}{2\pi} \int_0^{2\pi} f(t)e^{-int}\,dt = \frac{1}{2\pi} \sum_{j=-N}^{N} a_j \int_0^{2\pi} e^{i(j-n)t}\,dt\,.$$

Now the integral equals 0 if $j \neq n$ (this is so because $\int_0^{2\pi} e^{ikt}\,dt = 0$ when k is a nonzero integer). And the term with $j = n$ gives rise to $a_n \cdot 1$. Thus we find that

$$a_n = \frac{1}{2\pi} \int_0^{2\pi} f(t)e^{-int}\,dt\,. \qquad (11.2.2)$$

Since, in Exercise **5** of Section 9.3, we showed that functions of the form (11.2.1) are dense in the continuous functions, we might hope that a formula like (11.2.2) will give a method for calculating the coefficients of a trigonometric expansion in considerable generality. In any event, this calculation helps to justify (after the fact) our formula for $\widehat{f}(n)$.

EXAMPLE 11.1 Let $f(x) = x$. Then

$$a_n = \frac{1}{2\pi} \int_0^{2\pi} te^{-int}\,dt\,.$$

This is easily calculated to equal

$$a_n = -\frac{1}{in}\,.$$

Therefore the Fourier expansion of f is

$$\sum_{n=-\infty}^{\infty} \frac{-1}{in}e^{int}\,. \qquad \square$$

The other theory that you know for decomposing a function into simple components is the theory of Taylor series. However, in order for a function to have a Taylor series it must be infinitely differentiable. Even then, as we have learned, the Taylor series of a function usually does not converge, and if it does converge its limit may not be the original function—see Section 9.2. The Fourier series of f converges to f under fairly mild hypotheses on f, and thus provides a useful tool in analysis.

The first result we shall prove about Fourier series gives a growth condition on the coefficients $\widehat{f}(n)$:

Proposition 11.2 (Bessel's Inequality) *If f^2 is integrable then*

$$\sum_{n=-N}^{N} |\widehat{f_n}|^2 \leq \int_0^{2\pi} |f(t)|^2\,dt.$$

Proof: Recall that $\overline{e^{ijt}} = e^{-ijt}$ and $|a|^2 = a \cdot \overline{a}$ for $a \in \mathbb{C}$. We calculate

$$\frac{1}{2\pi} \int_0^{2\pi} |f(t) - S_N f(t)|^2 \, dt$$

$$= \frac{1}{2\pi} \int_0^{2\pi} \left(f(t) - \sum_{n=-N}^{N} \widehat{f}(n) e^{int} \right) \cdot \overline{\left(f(t) - \sum_{n=-N}^{N} \widehat{f}(n) e^{int} \right)} \, dt$$

$$= \frac{1}{2\pi} \int_0^{2\pi} |f(t)|^2 \, dt - \sum_{n=-N}^{N} \frac{1}{2\pi} \int_0^{2\pi} f(t) e^{-int} \, dt \cdot \overline{\widehat{f}(n)}$$

$$- \sum_{n=-N}^{N} \frac{1}{2\pi} \int_0^{2\pi} \overline{f(t) e^{-int}} \, dt \cdot \widehat{f}(n)$$

$$+ \sum_{m,n} \widehat{f}(m) \overline{\widehat{f}(n)} \frac{1}{2\pi} \int_0^{2\pi} e^{imt} \cdot e^{-int} \, dt \, .$$

Now each of the first two sums equals $\sum_{n=-N}^{N} |\widehat{f}(n)|^2$. In the last sum, any summand with $m \neq n$ equals 0. The summands with $m = n$ equal $|\widehat{f}(n)|^2$. Thus our equation simplifies to

$$\frac{1}{2\pi} \int_0^{2\pi} |f(t) - S_N f(t)|^2 \, dt \;=\; \frac{1}{2\pi} \int_0^{2\pi} |f(t)|^2 \, dt - \sum_{n=-N}^{N} |\widehat{f}(n)|^2 \, .$$

Since the left side is nonnegative, it follows that

$$\sum_{n=-N}^{N} |\widehat{f}(n)|^2 \leq \frac{1}{2\pi} \int_0^{2\pi} |f(t)|^2 \, dt \, ,$$

as desired. $\qquad \square$

Corollary 11.3 If f^2 is integrable then the Fourier coefficients $\widehat{f}(n)$ satisfy

$$\widehat{f}(n) \to 0 \quad \text{as} \quad n \to \infty \, .$$

Proof: Since $\sum |\widehat{f}(n)|^2 < \infty$ we know that $|\widehat{f}(n)|^2 \to 0$. This implies the result. $\square$

Remark 11.4 In fact, with a little extra effort, one can show that the conclusion of the corollary holds if only f is integrable. This entire matter is addressed from a slightly different point of view in Proposition 11.16 below.

Since the coefficients of the Fourier series, at least for a square integrable function, tend to zero, we might hope that the Fourier series will converge in some sense. Of course the best circumstance would be that $S_N f \to f$ (pointwise, or in some other manner). We now turn our attention this problem.

Proposition 11.5 (The Dirichlet Kernel) *If f is integrable then*

$$S_N f(x) = \frac{1}{2\pi} \int_0^{2\pi} D_N(x - t) f(t) \, dt,$$

where

$$D_N(t) = \frac{\sin(N + \frac{1}{2})t}{\sin \frac{1}{2}t}.$$

Proof: Observe that

$$
\begin{aligned}
S_N f(x) &= \sum_{n=-N}^{N} \widehat{f}(n) e^{inx} \\
&= \sum_{n=-N}^{N} \frac{1}{2\pi} \int_0^{2\pi} f(t) e^{-int} \, dt \cdot e^{inx} \\
&= \sum_{n=-N}^{N} \frac{1}{2\pi} \int_0^{2\pi} f(t) e^{in(x-t)} \, dt \\
&= \frac{1}{2\pi} \int_0^{2\pi} f(t) \left[\sum_{n=-N}^{N} e^{in(x-t)} \right] dt.
\end{aligned}
$$

Thus we are finished if we can show that the sum in [ ] equals $D_N(x - t)$.

Rewrite the sum as

$$\sum_{n=0}^{N} \left(e^{i(x-t)} \right)^n + \sum_{n=0}^{N} \left(e^{-i(x-t)} \right)^n - 1.$$

Then each of these last two sums is the partial sum of a geometric series. Thus we use the formula from Proposition 3.15 to write the last line as

$$\frac{e^{i(x-t)(N+1)} - 1}{e^{i(x-t)} - 1} + \frac{e^{-i(x-t)(N+1)} - 1}{e^{-i(x-t)} - 1} - 1.$$

We put everything over a common denominator to obtain

$$\frac{\cos N(x - t) - \cos(N + 1)(x - t)}{1 - \cos(x - t)}.$$

We write

$$N(x - t) = \left((N + \tfrac{1}{2})(x - t) - \tfrac{1}{2}(x - t) \right),$$

$$(N + 1)(x - t) = \left((N + \tfrac{1}{2})(x - t) + \tfrac{1}{2}(x - t) \right),$$

$$(x - t) = \tfrac{1}{2}(x - t) + \tfrac{1}{2}(x - t)$$

and use the sum formula for the cosine function to find that the last line equals

$$\frac{2 \sin\left((N + \tfrac{1}{2})(x - t) \right) \sin\left(\tfrac{1}{2}(x - t) \right)}{2 \sin^2\left(\tfrac{1}{2}(x - t) \right)}$$

$$= \frac{\sin(N + \tfrac{1}{2})(x - t)}{\sin \tfrac{1}{2}(x - t)}$$

$$= D_N(x - t).$$

That is the desired conclusion. $\square$

Remark 11.6 We have presented this particular proof of the formula for D_N because it is the most natural. It is by no means the shortest. Another proof is explored in the exercises.

Note also that, by a change of variable, the formula for $S_N f$ presented in the proposition can also be written as

$$S_N f(x) = \frac{1}{2\pi} \int_0^{2\pi} D_N(t) f(x - t) \, dt$$

provided we adhere to the convention of doing all arithmetic modulo multiples of 2π.

Lemma 11.7 *For any N it holds that*

$$\frac{1}{2\pi} \int_0^{2\pi} D_N(t) \, dt = 1 \,.$$

Proof: It would be quite difficult to prove this property of D_N from the formula that we just derived. However, if we look at the proof of the proposition we notice that

$$D_N(t) = \sum_{n=-N}^{N} e^{int} \,.$$

Hence

$$\frac{1}{2\pi} \int_0^{2\pi} D_N(t)\, dt \;=\; \frac{1}{2\pi} \int_0^{2\pi} \sum_{n=-N}^{N} e^{int}\, dt$$

$$=\; \sum_{n=-N}^{N} \frac{1}{2\pi} \int_0^{2\pi} e^{int}\, dt$$

$$=\; 1$$

because any power of e^{it}, except the zeroeth power, integrates to zero. This completes the proof. □

Next we prove that, for a large class of functions, the Fourier series converges back to the function at every point.

Theorem 11.8 *Let f be a function on $[0, 2\pi]$ that satisfies a Lipschitz condition: there is a constant $C > 0$ such that if $s, t \in [0, 2\pi]$ then*

$$|f(s) - f(t)| \le C \cdot |s - t|. \tag{11.8.1}$$

[Note that at 0 and 2π this condition is required to hold modulo 2π—see the remarks at the beginning of the section.] Then, for every $x \in [0, 2\pi]$, it holds that

$$S_N f(x) \to f(x) \quad \text{as} \quad N \to \infty.$$

Indeed, the convergence is uniform in x.

Proof: Fix $x \in [0, 2\pi]$. We calculate that

$$|S_N f(x) - f(x)| \;=\; \left| \frac{1}{2\pi} \int_0^{2\pi} f(x - t) D_N(t)\, dt - f(x) \right|$$

$$=\; \left| \frac{1}{2\pi} \int_0^{2\pi} f(x - t) D_N(t)\, dt \right.$$

$$\left. -\frac{1}{2\pi} \int_0^{2\pi} f(x) D_N(t)\, dt \right|,$$

where we have made use of the lemma. It is convenient here to exploit periodicity and write our integrals as $\int_{-\pi}^{\pi}$ instead of $\int_0^{2\pi}$. Now we combine the integrals to write

$$|S_N f(x) - f(x)|$$

$$= \left| \frac{1}{2\pi} \int_{-\pi}^{\pi} [f(x-t) - f(x)] D_N(t) \, dt \right|$$

$$= \left| \frac{1}{2\pi} \int_{-\pi}^{\pi} \left[\frac{f(x-t) - f(x)}{\sin t/2} \right] \cdot \sin \left((N + \frac{1}{2})t \right) \, dt \right|$$

$$\leq \left| \int \left[\frac{f(x-t) - f(x)}{\sin t/2} \cdot \cos \frac{t}{2} \right] \sin Nt \, dt \right|$$

$$+ \left| \frac{1}{2\pi} \int_{-\pi}^{\pi} \left[\frac{f(x-t) - f(x)}{\sin t/2} \cdot \sin \frac{t}{2} \right] \cos Nt \, dt \right|$$

$$\leq \left| \frac{1}{2\pi} \int_{-\pi}^{\pi} h(t) \sin Nt \, dt \right| + \left| \frac{1}{2\pi} \int_{-\pi}^{\pi} k(t) \cos Nt \, dt \right| ,$$

where we have denoted the first expression in [ ] by $h_x(t) = h(t)$ and the second expression in [ ] by $k_x(t) = k(t)$. We use our hypothesis (11.8.1) about f to see that

$$|h(t)| = \left| \frac{f(x-t) - f(x)}{t} \right| \cdot \left| \frac{t}{\sin(t/2)} \right| \cdot \left| \cos \frac{t}{2} \right| \leq C \cdot 3.$$

[Here we have used the elementary fact that $2/\pi \leq |\sin u/u| \leq 1$ for $-\pi/2 \leq u \leq \pi/2$.] Thus h is a bounded function. It is obviously continuous, because f is, except perhaps at $t = 0$. So h is integrable—since it is bounded it is even square integrable. An even easier discussion shows that k is square integrable. Therefore Corollary 11.3 applies and we may conclude that the Fourier coefficients of h and of k tend to zero. However, the integral involving h is nothing other than $(\widehat{h}(N) - \widehat{h}(-N))/(2i)$ and the integral involving k is precisely $(\widehat{k}(N) + \widehat{k}(-N))/2$. We conclude that these integrals tend to zero as $N \to \infty$; in other words,

$$|S_N f(x) - f(x)| \to 0 \quad \text{as} \quad N \to \infty.$$

Since the relevant estimates are independent of x, we see that the convergence is uniform. □

Corollary 11.9 If $f \in C^1([0, 2\pi])$ (that is, f is continuously differentiable) then $S_N f \to f$ uniformly.

Proof: A C^1 function, by the Mean Value Theorem, satisfies a Lipschitz condition. □

In fact the proof of the theorem suffices to show that, if f is a Riemann square-integrable function on $[0, 2\pi]$ and if f is differentiable at x, then $S_N f(x) \to f(x)$.

In the exercises we shall explore other methods of summing Fourier series that allow us to realize even discontinuous functions as the limits of certain Fourier expressions.

It is natural to ask whether the Fourier series of a function characterizes that function. We can now give a partial answer to this question:

Corollary 11.10 *If f is a function on $[0, 2\pi]$ that satisfies a Lipschitz condition and if the Fourier series of f is identically zero then $f \equiv 0$.*

Proof: By the preceding corollary, the Fourier series converges uniformly to f. But the Fourier series is 0. □

Corollary 11.11 *If f and g are functions on $[0, 2\pi]$ that satisfy a Lipschitz condition and if the Fourier coefficients of f are the same as the Fourier coefficients of g then $f \equiv g$.*

Proof: Apply the preceding corollary to $f - g$. □

EXAMPLE 11.12 Let $f(t) = t^2 - 2\pi t, 0 \le t \le 2\pi$. Then $f(0) = f(2\pi) = 0$ and f is Lipschitz modulo 2π. Calculating the Fourier series of f, setting $t = 0$, and using the theorem reveals that

$$\sum_{j=1}^{\infty} \frac{1}{j^2} = \frac{\pi^2}{6}.$$

You are requested to provide the details. □

Exercises

1. Find the Fourier series for the function
$$f(x) = \begin{cases} 0 & \text{if} \quad -\pi \le x < 0 \\ 1 & \text{if} \quad 0 \le x \le \frac{\pi}{2} \\ 0 & \text{if} \quad \frac{\pi}{2} < x \le \pi. \end{cases}$$

2. Find the Fourier series of the function
$$f(x) = \begin{cases} 0 & \text{if} \quad -\pi \le x < 0 \\ \sin x & \text{if} \quad 0 \le x \le \pi \end{cases}$$

3. Find the Fourier series for each of these functions. Pay special attention to the reasoning used to establish your conclusions; consider alternative lines of thought.

   (a) $f(x) = \pi$, $-\pi \le x \le \pi$
   (b) $f(x) = \sin x$, $-\pi \le x \le \pi$
   (c) $f(x) = \cos x$, $-\pi \le x \le \pi$

    (d) $f(x) = \pi + \sin x + \cos x$, $\quad -\pi \le x \le \pi$

4. Find the Fourier series for the function given by

    (a)

$$f(x) = \begin{cases} -a & \text{if} \quad -\pi \le x < 0 \\ a & \text{if} \quad 0 \le x \le \pi \end{cases}$$

    for a a positive real number.

    (b)

$$f(x) = \begin{cases} -1 & \text{if} \quad -\pi \le x < 0 \\ 1 & \text{if} \quad 0 \le x \le \pi \end{cases}$$

    (c)

$$f(x) = \begin{cases} -\frac{\pi}{4} & \text{if} \quad -\pi \le x < 0 \\ \frac{\pi}{4} & \text{if} \quad 0 \le x \le \pi \end{cases}$$

    (d)

$$f(x) = \begin{cases} -1 & \text{if} \quad -\pi \le x < 0 \\ 2 & \text{if} \quad 0 \le x \le \pi \end{cases}$$

    (e)

$$f(x) = \begin{cases} 1 & \text{if} \quad -\pi \le x < 0 \\ 2 & \text{if} \quad 0 \le x \le \pi \end{cases}$$

5. The functions $\sin^2 x$ and $\cos^2 x$ are both even. Show, without using any calculations, that the identities

$$\sin^2 x = \frac{1}{2}(1 - \cos 2x) = \frac{1}{2} - \frac{1}{2}\cos 2x$$

and

$$\cos^2 x = \frac{1}{2}(1 + \cos 2x) = \frac{1}{2} + \frac{1}{2}\cos 2x$$

are actually the Fourier series expansions of these functions.

6. Prove the trigonometric identities

$$\sin^3 x = \frac{3}{4}\sin x - \frac{1}{4}\sin 3x \quad \text{and} \quad \cos^3 x = \frac{3}{4}x + \frac{1}{4}\cos 3x$$

and show briefly, without calculation, that these are the Fourier series expansions of the functions $\sin^3 x$ and $\cos^3 x$.

7. Give another proof for the formula for $D_N(t)$ by completing the following outline:

    (a) $D_N(t) = \sum_{n=-N}^{N} e^{int}$;

    (b) $(e^{it} - 1) \cdot D_N(t) = e^{i(N+1)t} - e^{-iNt}$;

    (c) Multiply both sides of the last equation by $e^{-it/2}$.

(d) Conclude that $D_N(t) = \frac{\sin(N+1/2)t}{\sin(t/2)}$.

8. Complete the details of Example 11.12.

*   9. If f is integrable on the interval $[0, 2\pi]$ and if N is a nonnegative integer then define

$$\sigma_N f(x) = \frac{1}{N+1} \sum_{n=0}^{N} S_N f(x)\,.$$

This is called the Nth *Cesaro mean* for the Fourier series of f. Prove that

$$\sigma_N f(x) = \frac{1}{2\pi} \int_0^{2\pi} K_N(x-t) f(t)\, dt\,,$$

where

$$K_N(x-t) = \frac{1}{N+1} \left\{ \frac{\sin \frac{N+1}{2}(x-t)}{\sin \frac{1}{2} t} \right\}^2\,.$$

10. Refer to Exercise **13** for notation. Prove that if $\delta > 0$ then $\lim_{N \to \infty} K_N(t) = 0$ with the limit being uniform for all $|t| \geq \delta$.

*   11. Refer to Exercise **13** for notation. Prove that $\frac{1}{2\pi} \int_0^{2\pi} |K_N(t)|\, dt = 1$.

11.3   An Introduction to the Fourier Transform

It turns out that Fourier analysis on the interval $[0, 2\pi]$ and Fourier analysis on the entire real line $\mathbb{R}$ are analogous; but they differ in certain particulars that are well worth recording. In the present section we present an outline of the theory of the Fourier transform on the line. A thorough treatment of Fourier analysis in Euclidean space may be found in [STG]. See also [KRA2].

We define the *Fourier transform* of an integrable function f on $\mathbb{R}$ by

$$\widehat{f}(\xi) = \int_{\mathbb{R}} f(t) e^{it \cdot \xi}\, dt\,.$$

Many references will insert a factor of 2π in the exponential or in the measure. Others will insert a minus sign in the exponent. There is no agreement on this matter. We have opted for this particular definition because of its simplicity.

We note that the significance of the exponentials $e^{it \cdot \xi}$ is that the only continuous multiplicative homomorphisms of $\mathbb{R}$ into the circle group are the functions $\phi_\xi(t) = e^{it \cdot \xi}, \xi \in \mathbb{R}$. These functions are called the *characters* of the additive group $\mathbb{R}$. We refer the reader to [KRA2] for more on this matter.

Proposition 11.13 *If f is an integrable function, then*

$$|\widehat{f}(\xi)| \leq \int_{\mathbb{R}} |f(t)|\, dt\,.$$

Proof: Observe that, for any $\xi \in \mathbb{R}$,

$$|\widehat{f}(\xi)| = \left| \int_{\mathbb{R}} f(t)e^{it \cdot \xi}\, dt \right| \leq \int_{\mathbb{R}} |f(t)e^{it \cdot \xi}|\, dt \leq \int |f(t)|\, dt. \qquad \square$$

Proposition 11.14 *If f is integrable, f is differentiable, and f' is integrable, then*

$$(f')\widehat{}(\xi) = -i\xi\widehat{f}(\xi).$$

Proof: Integrate by parts: if f is an infinitely differentiable function that vanishes outside a compact set, then

$$
\begin{aligned}
(f')\widehat{}(\xi) &= \int f'(t)e^{it \cdot \xi}\, dt\, dt \\
&= -\int f(t)[e^{it \cdot \xi}]'\, dt \\
&= -i\xi \int f(t)e^{it \cdot \xi}\, dt \\
&= -i\xi\widehat{f}(\xi).
\end{aligned}
$$

[Of course the "boundary terms" in the integration by parts vanish since f vanishes outside a compact set.] The general case follows from a limiting argument (see the Appendix at the end of this section). $\qquad \square$

Proposition 11.15 *If f is integrable and ixf is integrable, then*

$$(itf)\widehat{} = \frac{d}{d\xi}\widehat{f}.$$

Proof: Differentiate under the integral sign:

$$
\begin{aligned}
\frac{d}{d\xi}\widehat{f}(\xi) &= \frac{d}{d\xi} \int_{\mathbb{R}} f(t)e^{it\xi}\, dt \\
&= \int_{\mathbb{R}} f(t)\frac{d}{d\xi}(e^{it\xi})\, dt \\
&= \int_{\mathbb{R}} f(t)ite^{it\xi}\, dt \\
&= (itf)\widehat{}.
\end{aligned}
$$
$\qquad \square$

Proposition 11.16 (The Riemann–Lebesgue Lemma) *If f is integrable, then*

$$\lim_{\xi \to \infty} |\widehat{f}(\xi)| = 0.$$

Proof: First assume that $g \in C^2(\mathbb{R})$ and vanishes outside a compact set. We know that $|\widehat{g}|$ is bounded. Also

$$\left|\xi^2 \widehat{g}(\xi)\right| = \left|[g'']\widehat{\ }\,\right| \leq \int_{\mathbb{R}} |g''(x)|\, dx = C'\,.$$

Then $(1 + |\xi|^2)\widehat{g}$ is bounded. Thus

$$|\widehat{g}(\xi)| \leq \frac{C''}{1 + |\xi|^2} \overset{|\xi| \to \infty}{\longrightarrow} 0\,.$$

This proves the result for $g \in C_c^2$. [Notice that the argument also shows that, if $g \in C^2(\mathbb{R})$ and vanishes outside a compact set, then $\widehat{g}$ is integrable.]

Now let f be an arbitrary integrable function. Then there is a function $\psi \in C^2(\mathbb{R})$, vanishing outside a compact set, such that

$$\int_{\mathbb{R}} |f(x) - \psi(x)|\, dx < \epsilon/2\,.$$

[See the Appendix to this section for the details of this assertion.] Choose M so large that, when $|\xi| > M$, then $|\widehat{\psi}(\xi)| < \epsilon/2$. Then, for $|\xi| > M$, we have

$$
\begin{aligned}
|\widehat{f}(\xi)| &= |(f - \psi)\widehat{\ }(\xi) + \widehat{\psi}(\xi)| \\
&\leq |(f - \psi)\widehat{\ }(\xi)| + |\widehat{\psi}(\xi)| \\
&\leq \int_{\mathbb{R}} |f(x) - \psi(x)|\, dx + \frac{\epsilon}{2} \\
&< \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon\,.
\end{aligned}
$$

This proves the result. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\square$

EXAMPLE 11.17 The Riemann–Lebesgue lemma is intuitively clear when viewed in the following way. Fix an integrable function f. An integrable function is well-approximated by a continuous function, so we may as well suppose that f is continuous. But a continuous function is well-approximated by a smooth function (see the Appendix to this section), so we may as well suppose that f is smooth. On a small interval I—say of length $1/M$—a smooth function is nearly constant. So, if we let $|\xi| \gg 2\pi M^2$, then the character $e^{i\xi \cdot x}$ will oscillate at least M times on I, and will therefore integrate against a constant to a value that is very nearly zero. As M becomes larger, this statement becomes more and more accurate. That is the Riemann–Lebesgue lemma. $\qquad\square$

Proposition 11.18 *Let f be integrable on $\mathbb{R}$. Then $\widehat{f}$ is uniformly continuous.*

Proof: Let us first assume that f is continuous and vanishes outside a compact set. Then

$$\lim_{\xi \to \xi_0} \widehat{f}(\xi) = \lim_{\xi \to \xi_0} \int f(x) e^{ix \cdot \xi}\, dx = \int \lim_{\xi \to \xi_0} f(x) e^{ix \cdot \xi}\, dx = \widehat{f}(\xi_0)\,.$$

[*Exercise:* Justify passing the limit under the integral sign.] Since $\widehat{f}$ also vanishes at ∞, the result is immediate when f is continuous and vanishing outside a compact set. The general result follows from an approximation argument (see the Appendix to this section). □

Let $C_0(\mathbb{R})$ denote the continuous functions on $\mathbb{R}$ that vanish at ∞. Equip this space with the supremum norm. Then our results show that the Fourier transform maps the integrable functions to C_0 continuously.

It is natural to ask whether the Fourier transform is univalent; put in other words, can we recover a function from its Fourier transform? If so, can we do so with an explicit integral formula? The answer to all these questions is "yes," but advanced techniques are required for the proofs. We cannot treat them here, but see [KRA2] for the details. We content ourselves with the formulation of a single result and its consequences.

Theorem 11.19 *Let f be a continuous, integrable function on $\mathbb{R}$ and suppose also that $\widehat{f}$ is integrable. Then*

$$f(x) = \frac{1}{2\pi} \int_{\mathbb{R}} \widehat{f}(\xi) e^{-ix\cdot\xi} \, d\xi$$

for every x.

Corollary 11.20 *If f is continuous and integrable and $\widehat{f}(\xi) \equiv 0$ then $f \equiv 0$.*

Corollary 11.21 *If f, g are continuous and integrable and $\widehat{f}(\xi) = \widehat{g}(\xi)$ then $f \equiv g$.*

We refer to the circle of ideas in this theorem and the two corollaries as "Fourier inversion." See [KRA2] for the details of all these assertions.

Appendix: Approximation by Smooth Functions

At several junctures in this section we have used the idea that an integrable function may be approximated by smooth functions. We take a moment now to discuss this notion. Not all of the details appear here, but the interested reader may supply them as an exercise.

Let f be any integrable function on the interval $[0, 1]$. Then f may be approximated by its Riemann sums in the following sense. Let

$$0 = x_0 < x_1 < \cdots < x_k = 1$$

be a partition of the interval. For $j = 1, \ldots, k$ define

$$h_j(x) = \begin{cases} 0 & \text{if} \quad 0 \le x < x_{j-1} \\ 1 & \text{if} \quad x_{j-1} \le x \le x_j \\ 0 & \text{if} \quad x_j < x \le 1. \end{cases}$$

Then the function

$$\mathcal{R}f(x) = \sum_{j=1}^{k} f(x_j) \cdot h_j(x)$$

is a piecewise constant approximation for f and the expression

$$\int_{\mathbb{R}} |f(x) - \mathcal{R}f(x)| \; dx \tag{11.22}$$

will be small if the mesh of the partition is sufficiently fine. In fact the expression (11.22) is a standard "distance between functions" that is used in mathematical analysis. We often denote this quantity by $\|f - \mathcal{R}f\|_{L^1}$ and we call it "the L^1 norm" or "L^1 distance." More generally, we call the expression

$$\int_{\mathbb{R}} |g(x)| \; dx \equiv \|g\|_{L^1}$$

the L^1 norm of the function g.

Now our strategy is to approximate each of the functions h_j by a "smooth" function. Let $f(x) = 10x^3 - 15x^4 + 6x^5$. Notice that $f(0) = 0$, $f(1) = 1$, and both f' and f'' vanish at 0 and at 1.

The model for the sort of smooth function we are looking for is

$$\psi(x) = \begin{cases} 0 & \text{if} & x < -2 \\ f(x+2) & \text{if} & -2 \le x \le -1 \\ 1 & \text{if} & -1 < x < 1 \\ f(2-x) & \text{if} & 1 \le x \le 2 \\ 0 & \text{if} & 2 < x. \end{cases}$$

Refer to Figure 11.1. You may calculate that this function is twice continuously differentiable. It vanishes outside the interval $[-2, 2]$. And it is identically equal to 1 on the interval $[-1, 1]$.

More generally, we will consider the functions

$$\psi_\delta(x) = \begin{cases} 0 & \text{if} & x < -1 - \delta \\ f\left(\dfrac{x + (1+\delta)}{\delta}\right) & \text{if} & -1 - \delta \le x \le -1 \\ 1 & \text{if} & -1 < x < 1 \\ f\left(\dfrac{(1+\delta) - x}{\delta}\right) & \text{if} & 1 \le x \le 1 + \delta \\ 0 & \text{if} & 1 + \delta < x. \end{cases}$$
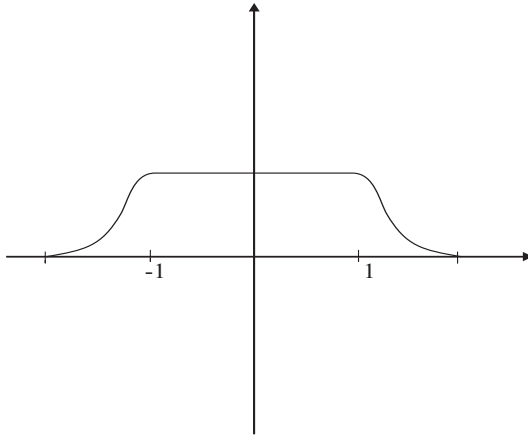
Figure 11.1: A compactly supported, smooth function.

for $\delta > 0$ and

$$\psi_\delta^{[a,b]}(x) = \psi_\delta \left(\frac{2x - b - a}{b - a} \right)$$

for $\delta > 0$ and $a < b$. Figure 11.2 shows that ψ_δ is similar to the function ψ, but its sides are contracted so that it climbs from 0 to 1 over the interval $[-1-\delta, -1]$ of length δ and then descends from 1 to 0 over the interval $[1, 1+\delta]$ of length δ. The function $\psi_\delta^{[a,b]}$ is simply the function ψ_δ adapted to the interval $[a, b]$ (Figure 11.3). The function $\psi_\delta^{[a,b]}$ climbs from 0 to 1 over the interval $[a - (\delta(b-a))/2, a]$ of length $\delta(b-a)/2$ and descends from 1 to 0 over the interval $[b, b+(\delta(b-a)/2)]$ of length $\delta(b - a)/2$.

Finally, we approximate the function h_j by $k_j(x) \equiv \psi_\delta^{[x_{j-1}, x_j]}$ for $j = 1, \ldots, k$. See Figure 11.4. Then the function f is approximated in L^1 norm by

$$\mathcal{S}f(x) = \sum_{j=1}^{k} f(x_j) \cdot k_j(x).$$

See Figure 11.5. If $\delta > 0$ is sufficiently small, then we can make $\|f - \mathcal{S}f\|_{L^1}$ as small as we please.

The approximation by twice continuously differentiable (or C^2) functions that we have constructed here is easily modified to achieve approximation by C^k functions for any k. One merely replaces the polynomial f by a polynomial that vanishes to higher order (order at least k) at 0 and at 1.

Exercises

1. Determine whether each of the following functions is even, odd, or neither:

$$x^5 \sin x , \quad x^2 \sin 2x , \quad e^x , \quad (\sin x)^3 , \quad \sin x^2 ,$$
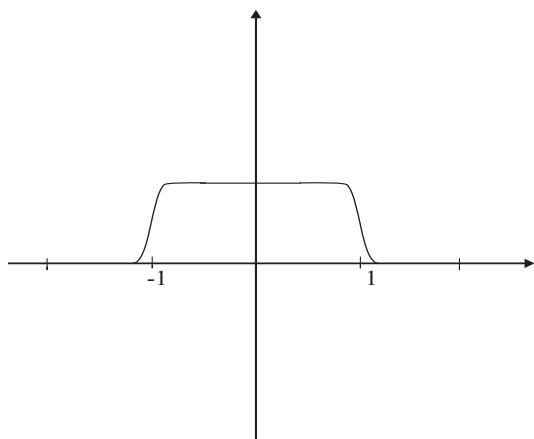
Figure 11.2: Another compactly supported, smooth function.
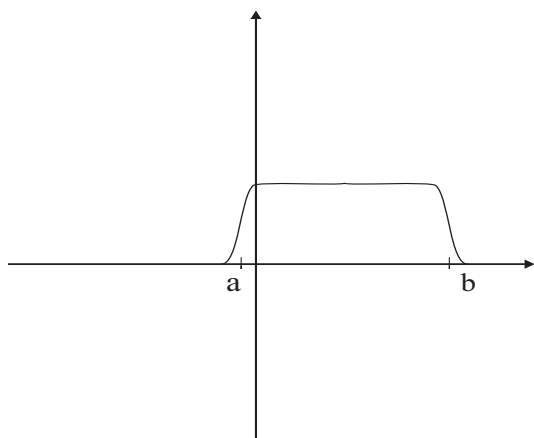
Figure 11.3: The compactly supported, smooth function translated and dilated.
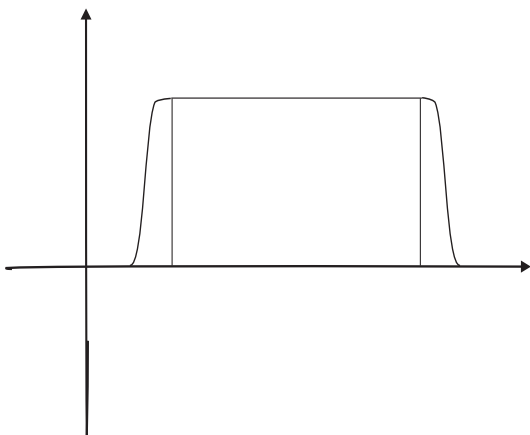
Figure 11.4: Unit for approximation.
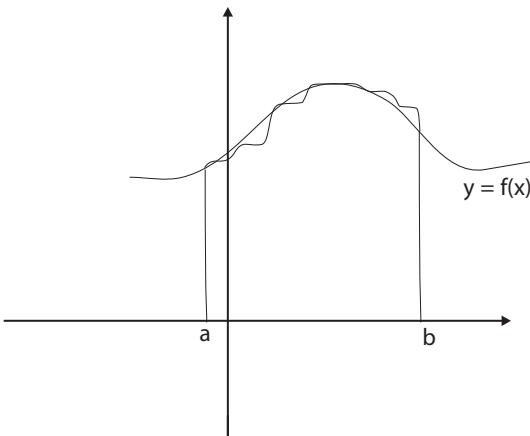
Figure 11.5: Approximation by a smooth function.

$$\cos(x + x^3) \ , \ x + x^2 + x^3 \ , \ \ln \frac{1+x}{1-x} \, .$$

2. Show that any function f defined on an interval symmetric about the origin can be written as the sum of an even function and an odd function.

3. Calculate the Fourier transform of $f(x) = x \cdot \chi_{[0,1]}$, where $\chi_{[0,1]}(x)$ equals 1 if $x \in [0, 1]$ and equals 0 otherwise.

4. See Exercise **4** for notation. Calculate the Fourier transform of $g(x) = \cos x \cdot \chi_{[0,2]}$.

5. If f, g are integrable functions on $\mathbb{R}$, then define their *convolution* to be

$$h(x) = f * g(x) = \int_{\mathbb{R}} f(x - t)g(t) \, dt \, .$$

Prove that

$$\widehat{h}(\xi) = \widehat{f}(\xi) \cdot \widehat{g}(\xi) \, .$$

6. Refer to Exercise **6** for notation and terminology. Fix an integrable function g on $\mathbb{R}$. Define a linear operator by

$$T : f \to f * g \, .$$

Prove that

$$\|Tf\|_{L^1} \le C\|f\|_{L^1} \, ,$$

where

$$\|f\|_{L^1} = \int_{\mathbb{R}} |f(x)| \, dx$$

and $C = \int |g| \, dx = \|g\|_{L^1}$.

* 7. Let f be a function on $\mathbb{R}$ that vanishes outside a compact set. Prove that $\widehat{f}$ does *not* vanish outside any compact set.

* 8. Calculate the Fourier transform of the function $f(x) = e^{-x^2}$.

* 9. Use the calculation from Exercise **9** to discover an eigenfunction of the Fourier transform.

* 10. Refer to Exercise **9**. What are the eigenvalues of the Fourier transform?

* 11. A version of the Poisson summation formula says that, if f is a suitable function on the real line, then

$$\sum_{n=-\infty}^{\infty} f(n) = \sum_{k=-\infty}^{\infty} \widehat{f}(k) \, .$$

Find a proof of this assertion.

11.4   Fourier Methods in the Theory of Differential Equations

In fact an entire separate book could be written about the applications of Fourier analysis to differential equations and to other parts of mathematical analysis. The subject of Fourier series grew up hand in hand with the analytical areas to which it is applied. In the present brief section we merely indicate a couple of examples.

11.4.1   Remarks on Different Fourier Notations

In Section 11.2, we found it convenient to define the Fourier coefficients of an integrable function on the interval $[0, 2\pi]$ to be

$$\widehat{f}(n) = \frac{1}{2\pi} \int_0^{2\pi} f(x) e^{-inx} \, dx \, .$$

From the point of view of pure mathematics, this complex notation has proved to be useful, and it has become standardized.

But, in applications, there are other Fourier paradigms. They are easily seen to be equivalent to the one we have already introduced. The reader who wants to be conversant in this subject should be aware of these different ways of writing the basic ideas of Fourier series. We will introduce one of them now, and use it in the ensuing discussion.

If f is integrable on the interval $[-\pi, \pi]$ (note that, by 2π-periodicity, this is not essentially different from $[0, 2\pi]$), then we define the Fourier coefficients

$$a_0 = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) \, dx \, ,$$

$$a_n = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \cos nx \, dx \qquad \text{for } n \geq 1 \, ,$$

$$b_n = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \sin nx \, dx \qquad \text{for } n \geq 1 \, .$$

This new notation is not essentially different from the old, for

$$\widehat{f}(n) = \frac{1}{2} \big[a_n + i b_n \big]$$

for $n \geq 1$. The change in normalization (i.e., whether the constant before the integral is $1/\pi$ or $1/2\pi$) is dictated by the observation that we want to exploit the fact (so that our formulas come out in a neat and elegant fashion) that

$$\frac{1}{2\pi} \int_0^{2\pi} |e^{-int}|^2 \, dt = 1 \, ,$$

in the theory from Section 11.2 and that

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} 1^2 \, dx = 1 \,,$$

$$\frac{1}{\pi} \int_{-\pi}^{\pi} |\cos nt|^2 \, dt = 1 \quad \text{for } n \geq 1 \,,$$

$$\frac{1}{\pi} \int_{-\pi}^{\pi} |\sin nt|^2 \, dt = 1 \quad \text{for } n \geq 1$$

in the theory that we are about to develop.

It is clear that any statement (as in Section 11.2) that is formulated in the language of $\widehat{f}(n)$ is easily translated into the language of a_n and b_n and vice versa. In the present discussion we shall use a_n and b_n just because that is the custom, and because it is convenient for the points that we want to make.

11.4.2   The Dirichlet Problem on the Disc

We now study the two-dimensional Laplace equation, which is

$$\triangle = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0 \,. \tag{11.4.1}$$

This is probably the most important differential equation of mathematical physics. It describes a steady-state heat distribution, electrical fields, and many other important phenomena of nature.

It will be useful for us to write this equation in polar coordinates. To do so, recall that

$$r^2 = x^2 + y^2 \,, \quad x = r\cos\theta \,, \quad y = r\sin\theta \,.$$

Thus

$$\frac{\partial}{\partial r} = \frac{\partial x}{\partial r}\frac{\partial}{\partial x} + \frac{\partial y}{\partial r}\frac{\partial}{\partial y} = \cos\theta \frac{\partial}{\partial x} + \sin\theta \frac{\partial}{\partial y}$$

$$\frac{\partial}{\partial \theta} = \frac{\partial x}{\partial \theta}\frac{\partial}{\partial x} + \frac{\partial y}{\partial \theta}\frac{\partial}{\partial y} = -r\sin\theta \frac{\partial}{\partial x} + r\cos\theta \frac{\partial}{\partial y}$$

We may solve these two equations for the unknowns $\partial/\partial x$ and $\partial/\partial y$. The result is

$$\frac{\partial}{\partial x} = \cos\theta \frac{\partial}{\partial r} - \frac{\sin\theta}{r}\frac{\partial}{\partial \theta} \quad \text{and} \quad \frac{\partial}{\partial y} = \sin\theta \frac{\partial}{\partial r} - \frac{\cos\theta}{r}\frac{\partial}{\partial \theta} \,.$$

A tedious calculation now reveals that

$$\begin{aligned}
\triangle = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} &= \left(\cos\theta \frac{\partial}{\partial r} - \frac{\sin\theta}{r}\frac{\partial}{\partial \theta} \right)\left(\cos\theta \frac{\partial}{\partial r} - \frac{\sin\theta}{r}\frac{\partial}{\partial \theta} \right) \\
&\quad + \left(\sin\theta \frac{\partial}{\partial r} - \frac{\cos\theta}{r}\frac{\partial}{\partial \theta} \right)\left(\sin\theta \frac{\partial}{\partial r} - \frac{\cos\theta}{r}\frac{\partial}{\partial \theta} \right) \\
&= \frac{\partial^2}{\partial r^2} + \frac{1}{r}\frac{\partial}{\partial r} + \frac{1}{r^2}\frac{\partial^2}{\partial \theta^2} \,.
\end{aligned}$$

Let us use the so-called separation of variables method to analyze our partial differential equation (11.4.1). We will seek a solution $w = w(r, \theta) = u(r) \cdot v(\theta)$ of the Laplace equation. Using the polar form, we find that this leads to the equation

$$u''(r) \cdot v(\theta) + \frac{1}{r}u'(r) \cdot v(\theta) + \frac{1}{r^2}u(r) \cdot v''(\theta) = 0 \,.$$

Thus

$$\frac{r^2u''(r) + ru'(r)}{u(r)} = -\frac{v''(\theta)}{v(\theta)} \,.$$

Since the left-hand side depends only on r, and the right-hand side only on θ, both sides must be constant. Denote the common constant value by λ.

Then we have

$$v''(\theta) + \lambda v(\theta) = 0 \tag{11.4.2}$$

and

$$r^2u''(r) + ru'(r) - \lambda u(r) = 0 \,. \tag{11.4.3}$$

If we demand that v be continuous and periodic, then we must insist that $\lambda > 0$ and in fact that $\lambda = n^2$ for some nonnegative integer n.[2] For $n = 0$ the only suitable solution is $v \equiv$ constant and for $n > 0$ the general solution (with $\lambda = n^2$) is

$$v = A \cos n\theta + B \sin n\theta \,,$$

as you can verify directly.

We set $\lambda = n^2$ in equation (11.4.3), and obtain

$$r^2u'' + ru' - n^2u = 0 \,, \tag{11.4.4}$$

which is Euler's equidimensional equation. The change of variables $r = e^z$ transforms this equation to a linear equation with constant coefficients, and that can in turn be solved with standard techniques. To wit, the equation that we now have is

$$u'' - n^2u = 0 \,.$$

The variable is now z. We guess a solution of the form $u(z) = e^{\alpha z}$. Thus

$$\alpha^2 e^{\alpha z} - n^2 e^{\alpha z} = 0 \tag{11.4.5}$$

so that

$$\alpha = \pm n \,.$$

Hence the solutions of (11.4.5) are

$$u(z) = e^{nz} \quad \text{and} \quad u(z) = e^{-nz}$$

provided that $n \neq 0$. It follows that the solutions of the original Euler equation (11.4.4) are

$$u(r) = r^n \quad \text{and} \quad u(r) = r^{-n} \qquad \text{for } n \neq 0 \,.$$

---

[2]More explicitly, $\lambda = 0$ gives a linear function for a solution and $\lambda < 0$ gives an exponential function for a solution.

In case $n = 0$ the solution is readily seen to be $u = 1$ or $u = \ln r$.

The result is

$$u = A + B \ln r \qquad \text{if } n = 0 \, ;$$

$$u = Ar^n + Br^{-n} \qquad \text{if } n = 1, 2, 3, \ldots \, .$$

We are most interested in solutions u that are continuous at the origin; so we take $B = 0$ in all cases. The resulting solutions are

$$
\begin{aligned}
n &= 0 \, , & w &= \text{a constant } a_0/2 \, ; \\
n &= 1 \, , & w &= r(a_1 \cos \theta + b_1 \sin \theta) \, ; \\
n &= 2 \, , & w &= r^2(a_2 \cos 2\theta + b_2 \sin 2\theta) \, ; \\
n &= 3 \, , & w &= r^3(a_3 \cos 3\theta + b_3 \sin 3\theta) \, ; \\
& & & \cdots
\end{aligned}
$$

Of course any finite sum of solutions of Laplace's equation is also a solution. The same is true for infinite sums. Thus we are led to consider

$$w = w(r, \theta) = \frac{1}{2}a_0 + \sum_{j=0}^{\infty} r^j(a_j \cos j\theta + b_j \sin j\theta) \, .$$

On a formal level, letting $r \to 1^-$ in this last expression gives

$$w = \frac{1}{2}a_0 + \sum_{j=1}^{\infty}(a_j \cos j\theta + b_j \sin j\theta) \, .$$

We draw all these ideas together with the following physical rubric. Consider a thin aluminum disc of radius 1, and imagine applying a heat distribution to the boundary of that disc. In polar coordinates, this distribution is specified by a function $f(\theta)$. We seek to understand the steady-state heat distribution on the entire disc. See Figure 11.6. So we seek a function $w(r, \theta)$, continuous on the closure of the disc, which agrees with f on the boundary and which represents the steady-state distribution of heat inside. Some physical analysis shows that such a function w is the solution of the boundary value problem

$$
\begin{aligned}
\triangle w &= 0 \, , \\
u\Big|_{\partial D} &= f \, .
\end{aligned}
$$

According to the calculations we performed prior to this last paragraph, a natural approach to this problem is to expand the given function f in its sine/cosine series:

$$f(\theta) = \frac{1}{2}a_0 + \sum_{j=1}^{\infty}(a_j \cos j\theta + b_j \sin j\theta)$$
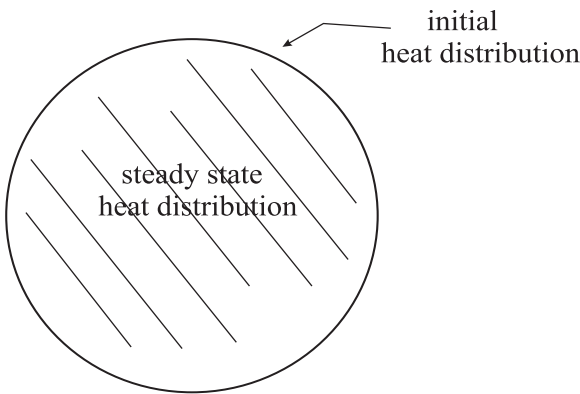
Figure 11.6: Steady-state heat.
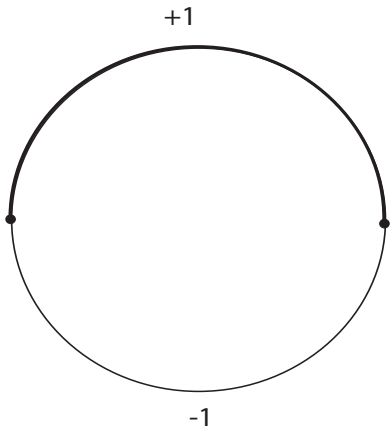
Figure 11.7: Boundary data.

and then posit that the w we seek is

$$w(r, \theta) = \frac{1}{2}a_0 + \sum_{j=1}^{\infty} r^j \left(a_j \cos j\theta + b_j \sin j\theta \right).$$

This process is known as *solving the Dirichlet problem on the disc with boundary data f*.

EXAMPLE 11.22 Let us follow the paradigm just sketched to solve the Dirichlet problem on the disc with $f(\theta) = 1$ on the top half of the boundary and $f(\theta) = -1$ on the bottom half of the boundary. See Figure 11.7.

It is straightforward to calculate that the Fourier series (sine series) expan-
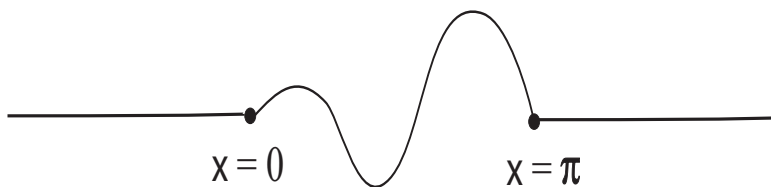
Figure 11.8: The wave equation.

sion for this f is

$$f(\theta) = \frac{4}{\pi}\left(\sin\theta + \frac{\sin 3\theta}{3} + +\frac{\sin 5\theta}{5} + \cdots\right).$$

There are no cosine terms because f is an odd function.

The solution of the Dirichlet problem is therefore

$$w(r,\theta) = \frac{4}{\pi}\left(r\sin\theta + \frac{r^3\sin 3\theta}{3} + +\frac{r^5\sin 5\theta}{5} + \cdots\right).$$

$\square$

11.4.3   Introduction to the Heat and Wave Equations

In the middle of the eighteenth century much attention was given to the problem of determining the mathematical laws governing the motion of a vibrating string with fixed endpoints at 0 and π (Figure 11.8). An elementary analysis of tension shows that, if $y(x,t)$ denotes the ordinate of the string at time t above the point x, then $y(x,t)$ satisfies the *wave equation*

$$\frac{\partial^2 y}{\partial t^2} = a^2 \frac{\partial^2 y}{\partial x^2}.$$

Here a is a parameter that depends on the tension of the string. A change of scale will allow us to assume that $a = 1$. (A bit later we shall actually provide a formal derivation of the wave equation. See also [KRA2] for a more thorough consideration of these matters.)

In 1747 d'Alembert showed that solutions of this equation have the form

$$y(x,t) = \frac{1}{2}\left(f(t+x) + g(t-x)\right), \tag{11.4.6}$$

where f and g are "any" functions of one variable. (The following technicality must be noted: the functions f and g are initially specified on the interval $[0,\pi]$. We extend f and g to $[-\pi, 0]$ and to $[\pi, 2\pi]$ by odd reflection. Continue f and g to the rest of the real line so that they are 2π-periodic.)

In fact the wave equation, when placed in a "well-posed" setting, comes equipped with two initial conditions:

$$\textbf{(i)} \qquad y(x,0) \;=\; \phi(x)$$
$$\textbf{(ii)} \qquad \partial_t y(x,0) \;=\; \psi(x).$$

These conditions mean **(i)** that the wave has an initial configuration that is the graph of the function ϕ and **(ii)** that the string is released with initial velocity ψ.

If (11.4.6) is to be a solution of this initial value problem then f and g must satisfy

$$\frac{1}{2}\left(f(x) + g(-x)\right) = \phi(x) \tag{11.4.7}$$

and

$$\frac{1}{2}\left(f'(x) + g'(-x)\right) = \psi(x). \tag{11.4.8}$$

Integration of (11.4.8) gives a formula for $f(x) - g(-x)$. That and (11.4.7) give a system that may be solved for f and g with elementary algebra.

The converse statement holds as well: for any functions f and g, a function y of the form (11.4.6) satisfies the wave equation (Exercise). The work of d'Alembert brought to the fore a controversy which had been implicit in the work of Daniel Bernoulli, Leonhard Euler, and others: what is a "function"? (We recommend the article [LUZ] for an authoritative discussion of the controversies that grew out of classical studies of the wave equation. See also [LAN].)

It is clear, for instance, in Euler's writings that he did not perceive a function to be an arbitrary "rule" that assigns points of the range to points of the domain; in particular, Euler did not think that a function could be specified in a fairly arbitrary fashion at different points of the domain. Once a function was specified on some small interval, Euler thought that it could only be extended in one way to a larger interval. Therefore, on physical grounds, Euler objected to d'Alembert's work. He claimed that the initial position of the vibrating string could be specified by several different functions pieced together continuously, so that a single f could not generate the motion of the string.

Daniel Bernoulli solved the wave equation by a different method (separation of variables, which we treat below) and was able to show that there are infinitely many solutions of the wave equation having the form

$$\phi_j(x,t) = \sin jx \cos jt\,, \quad j \geq 1 \text{ an integer}.$$

Proceeding formally, he posited that all solutions of the wave equation satisfying $y(0,t) = y(\pi, t) = 0$ and $\partial_t y(x,0) = 0$ will have the form

$$y = \sum_{j=1}^{\infty} a_j \sin jx \cos jt.$$

Setting $t = 0$ indicates that the initial form of the string is $f(x) \equiv \sum_{j=1}^{\infty} a_j \sin jx$. In d'Alembert's language, the initial form of the string is

$\frac{1}{2}\big(f(x) - f(-x)\big)$, for we know that

$$0 \equiv y(0, t) = f(t) + g(t)$$

(because the endpoints of the string are held stationary), hence $g(t) = -f(t)$. If we suppose that d'Alembert's function is odd (as is $\sin jx$, each j), then the initial position is given by $f(x)$. Thus the problem of reconciling Bernoulli's solution to d'Alembert's reduces to the question of whether an "arbitrary" function f on $[0, \pi]$ may be written in the form $\sum_{j=1}^{\infty} a_j \sin jx$.

Since most mathematicians contemporary with Bernoulli believed that properties such as continuity, differentiability, and periodicity were preserved under (even infinite) addition, the consensus was that arbitrary f could *not* be represented as a (even infinite) trigonometric sum. The controversy extended over some years and was fueled by further discoveries (such as Lagrange's technique for interpolation by trigonometric polynomials) and more speculations.

In the 1820s, the problem of representation of an "arbitrary" function by trigonometric series was given a satisfactory answer as a result of two events. First, there is the sequence of papers by Joseph Fourier culminating with the tract [FOU]. Fourier gave a formal method of expanding an "arbitrary" function f into a trigonometric series. He computed some partial sums for some sample fs and verified that they gave very good approximations to f. Second, Dirichlet proved the first theorem giving sufficient (and very general) conditions for the Fourier series of a function f to converge pointwise to f. *Dirichlet was one of the first, in 1828, to formalize the notions of partial sum and convergence of a series*; his ideas had antecedents in the work of Gauss and Cauchy.

For all practical purposes, these events mark the beginning of the mathematical theory of Fourier series (see [LAN]).

11.4.4 Boundary Value Problems

We wish to motivate the physics of the vibrating string. We begin this discussion by seeking a nontrivial solution y of the differential equation

$$y'' + \lambda y = 0 \tag{11.4.9}$$

subject to the conditions

$$y(0) = 0 \quad \text{and} \quad y(\pi) = 0. \tag{11.4.10}$$

Notice that this is a different situation from the one we have studied in earlier parts of the book. Ordinary differential equations generally have "initial conditions." Now we have what are called *boundary conditions*: we specify one condition (in this instance the *value*) for the function at two different points. For instance, in the discussion of the vibrating string in the last section, we wanted our string to be pinned down at the two endpoints. These are typical boundary conditions coming from a physical problem.

The situation with boundary conditions is quite different from that for initial conditions. The latter is a sophisticated variation of the fundamental theorem of calculus. The former is rather more subtle. So let us begin to analyze.

First, we can just solve the equation explicitly when $\lambda < 0$ and see that the independent solutions are a pair of exponentials, no linear combination of which can satisfy (11.4.10).

If $\lambda = 0$ then the general solution of (11.4.9) is the linear function $y = Ax + B$. Such a function cannot vanish at two points unless it is identically zero.

So the only interesting case is $\lambda > 0$. In this situation, the general solution of (11.4.9) is

$$y = A \sin \sqrt{\lambda} x + B \cos \sqrt{\lambda} x \, .$$

Since $y(0) = 0$, this in fact reduces to

$$y = A \sin \sqrt{\lambda} x \, .$$

In order for $y(\pi) = 0$, we must have $\sqrt{\lambda} \pi = n\pi$ for some positive integer n, thus $\lambda = n^2$. These values of λ are termed the *eigenvalues* of the problem, and the corresponding solutions

$$\sin x \, , \quad \sin 2x \, , \quad \sin 3x \, \ldots$$

are called the *eigenfunctions* of the problem (11.4.9), (11.4.10).

We note these immediate properties of the eigenvalues and eigenfunctions for our problem:

(i) If ϕ is an eigenfunction for eigenvalue λ, then so is $c \cdot \phi$ for any constant c.

(ii) The eigenvalues $1, 4, 9, \ldots$ form an increasing sequence that approaches $+\infty$.

(iii) The nth eigenfunction $\sin nx$ vanishes at the endpoints $0, \pi$ (as we originally mandated) and has exactly $n - 1$ zeros in the interval $(0, \pi)$.

11.4.5   Derivation of the Wave Equation

Now let us re-examine the vibrating string from the last section and see how eigenfunctions and eigenvalues arise naturally in a physical problem. We consider a flexible string with negligible weight that is fixed at its ends at the points $(0, 0)$ and $(\pi, 0)$. The curve is deformed into an initial position $y = f(x)$ in the x-y plane and then released.

Our analysis will ignore damping effects, such as air resistance. We assume that, in its relaxed position, the string is as in Figure 11.9. The string is plucked in the vertical direction, and is thus set in motion in a vertical plane. We will be supposing that the oscillation has small amplitude.

We focus attention on an "element" Δx of the string (Figure 11.10) that lies between x and $x + \Delta x$. We adopt the usual physical conceit of assuming that the displacement (motion) of this string element is *small*, so that there is only a slight error in supposing that the motion of each point of the string element is strictly vertical. We let the tension of the string, at the point x at

Figure 11.9: The string in relaxed position.
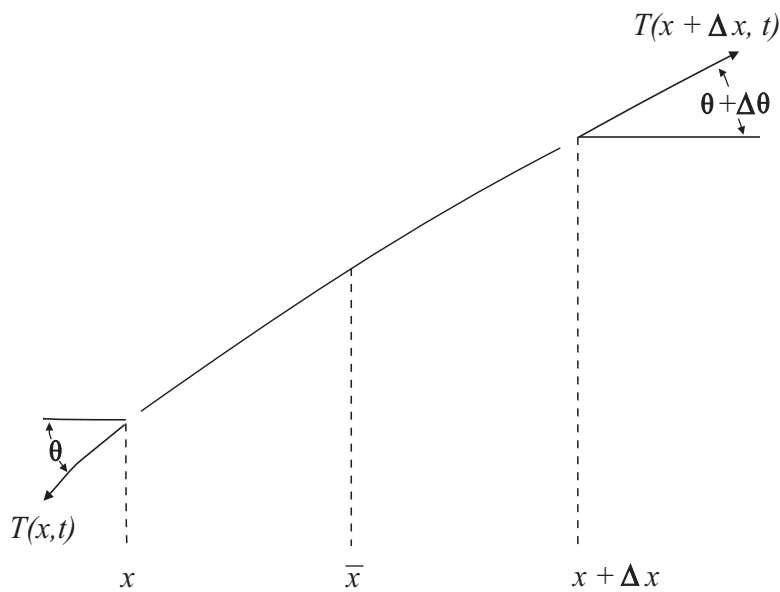
Figure 11.10: An element of the plucked string.
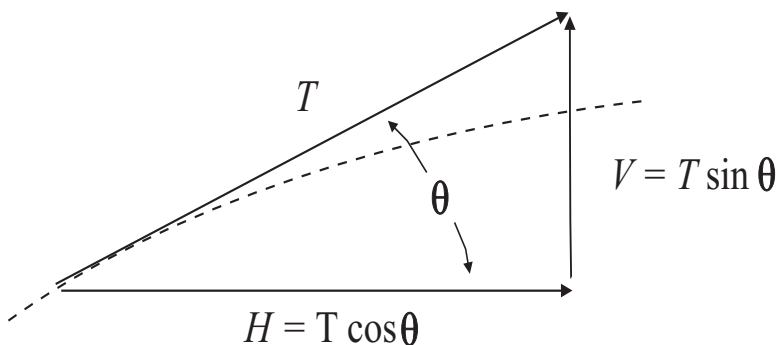
Figure 11.11: The horizontal component of the tension.

time t, be denoted by $T(x, t)$. Note that T acts only in the tangential direction (i.e., along the string). We denote the mass density of the string by ρ.

Since *there is no horizontal component of acceleration*, we see that

$$T(x + \Delta x, t) \cdot \cos(\theta + \Delta \theta) - T(x, t) \cdot \cos(\theta) = 0. \qquad (11.4.11)$$

(Refer to Figure 11.11: The expression $T(\star) \cdot \cos(\star)$ denotes $H(\star)$, the horizontal component of the tension.) Thus equation (6) says that H is independent of x.

Now we look at the vertical component of force (acceleration):

$$T(x + \Delta x, t) \cdot \sin(\theta + \Delta \theta) - T(x, t) \cdot \sin(\theta) = \rho \cdot \Delta x \cdot u_{tt}(\overline{x}, t). \qquad (11.4.12)$$

Here $\overline{x}$ is the mass center of the string element and we are applying Newton's second law—that the external force is the mass of the string element times the acceleration of its center of mass. We use subscripts to denote derivatives. We denote the vertical component of $T(\star)$ by $V(\star)$. Thus equation (11.4.12) can be written as

$$\frac{V(x + \Delta x, t) - V(x, t)}{\Delta x} = \rho \cdot u_{tt}(x, t).$$

Letting $\Delta x \to 0$ yields

$$V_x(x, t) = \rho \cdot u_{tt}(x, t). \qquad (11.4.13)$$

We would like to express equation (11.4.13) entirely in terms of u, so we notice that

$$V(x, t) = H(t) \tan \theta = H(t) \cdot u_x(x, t).$$

(We have used the fact that the derivative in x is the slope of the tangent line, which is $\tan \theta$.) Substituting this expression for V into (11.4.13) yields

$$(H u_x)_x = \rho \cdot u_{tt}.$$

But H is independent of x, so this last line simplifies to

$$H \cdot u_{xx} = \rho \cdot u_{tt}.$$

For small displacements of the string, θ is nearly zero, so $H = T \cos \theta$ is nearly T. We are most interested in the case where T is constant. And of course ρ is constant. Thus we finally write our equation as

$$\frac{T}{\rho} u_{xx} = u_{tt}.$$

It is traditional to denote the constant T/ρ on the left by a^2. We finally arrive at the *wave equation*

$$a^2 u_{xx} = u_{tt}.$$

11.4.6   Solution of the Wave Equation

We consider the wave equation

$$a^2 y_{xx} = y_{tt} \qquad\qquad (11.4.14)$$

with the boundary conditions

$$y(0, t) = 0$$

and

$$y(\pi, t) = 0 \,.$$

Physical considerations dictate that we also impose the initial conditions

$$\left. \frac{\partial y}{\partial t} \right|_{t=0} = 0 \qquad\qquad (11.4.15)$$

(indicating that the initial velocity of the string is 0) and

$$y(x, 0) = f(x) \qquad\qquad (11.4.16)$$

(indicating that the initial configuration of the string is the graph of the function f).

We solve the wave equation using a classical technique known as "separation of variables." For convenience, we assume that the constant $a = 1$. We guess a solution of the form $u(x, t) = u(x) \cdot v(t)$. Putting this guess into the differential equation

$$u_{xx} = u_{tt}$$

gives

$$u''(x)v(t) = u(x)v''(t) \,.$$

We may obviously separate variables, in the sense that we may write

$$\frac{u''(x)}{u(x)} = \frac{v''(t)}{v(t)} \,.$$

The left-hand side depends only on x while the right-hand side depends only on t. The only way this can be true is if

$$\frac{u''(x)}{u(x)} = \lambda = \frac{v''(t)}{v(t)}$$

for some constant λ. But this gives rise to two second-order linear, ordinary differential equations that we can solve explicitly:

$$u'' = \lambda \cdot u \qquad\qquad (11.4.17)$$

$$v'' = \lambda \cdot v\,. \qquad\qquad (11.4.18)$$

Observe that this is the *same* constant λ in both of these equations. Now, as we have already discussed, we want the initial configuration of the string to pass through the points $(0,0)$ and $(\pi, 0)$. We can achieve these conditions by solving (11.4.17) with $u(0) = 0$ and $u(\pi) = 0$. But of course this is the eigenvalue problem that we treated at the beginning of the section. The problem has a nontrivial solution if and only if $\lambda = -n^2$ for some positive integer n, and the corresponding eigenfunction is

$$u_n(x) = \sin nx\,.$$

For this same λ, the general solution of (11.4.15) is

$$v(t) = A\sin nt + B\cos nt\,.$$

If we impose the requirement that $v'(0) = 0$, so that (10) is satisfied, then $A = 0$ and we find the solution

$$v(t) = B\cos nt\,.$$

This means that the solution we have found of our differential equation with boundary and initial conditions is

$$y_n(x,t) = \sin nx \cos nt\,. \qquad\qquad (11.4.19)$$

And in fact any finite sum with coefficients (or *linear combination*) of these solutions will also be a solution:

$$y = \alpha_1 \sin x \cos t + \alpha_2 \sin 2x \cos 2t + \cdots \alpha_k \sin kx \cos kt\,.$$

Ignoring the rather delicate issue of convergence, we may claim that any *infinite* linear combination of the solutions (11.4.16) will also be a solution:

$$y = \sum_{j=1}^{\infty} b_j \sin jx \cos jt\,. \qquad\qquad (11.4.20)$$

Now we must examine the initial condition (11.4.16). The mandate $y(x,0) = f(x)$ translates to

$$\sum_{j=1}^{\infty} b_j \sin jx = y(x,0) = f(x) \qquad\qquad (11.4.21)$$

or

$$\sum_{j=1}^{\infty} b_j u_j(x) = y(x,0) = f(x). \qquad (11.4.22)$$

Thus we demand that f have a valid Fourier series expansion. Such an expansion is correct for a rather broad class of functions f. Thus the wave equation is solvable in considerable generality.

Now fix $m \neq n$. We know that our eigenfunctions u_j satisfy

$$u_m'' = -m^2 u_m \qquad \text{and} \qquad u_n'' = -n^2 u_n.$$

Multiply the first equation by u_n and the second by u_m and subtract. The result is

$$u_n u_m'' - u_m u_n'' = (n^2 - m^2) u_n u_m$$

or

$$[u_n u_m' - u_m u_n']' = (n^2 - m^2) u_n u_m.$$

We integrate both sides of this last equation from 0 to π and use the fact that $u_j(0) = u_j(\pi) = 0$ for every j. The result is

$$0 = [u_n u_m' - u_m u_n']\Big|_0^\pi = (n^2 - m^2) \int_0^\pi u_m(x) u_n(x)\, dx.$$

Thus

$$\int_0^\pi \sin mx \sin nx\, dx = 0 \qquad \text{for } n \neq m \qquad (11.4.23)$$

or

$$\int_0^\pi u_m(x) u_n(x)\, dx = 0 \qquad \text{for } n \neq m. \qquad (11.4.24)$$

Of course this is a standard fact from calculus. But now we understand it as an orthogonality condition, and we see how the condition arises naturally from the differential equation.

In view of the orthogonality condition (11.4.24), it is natural to integrate both sides of (11.4.22) against $u_k(x)$. The result is

$$
\begin{aligned}
\int_0^\pi f(x) \cdot u_k(x)\, dx &= \int_0^\pi \left(\sum_{j=0}^{\infty} b_j u_j(x) \right) \cdot u_k(x)\, dx \\
&= \sum_{j=0}^{\infty} b_j \int_0^\pi u_j(x) u_k(x)\, dx \\
&= \frac{\pi}{2} b_k.
\end{aligned}
$$

The b_k are the Fourier coefficients that we studied earlier in this chapter. Using these coefficients, we have *Bernoulli's solution* (11.4.20) of the wave equation.
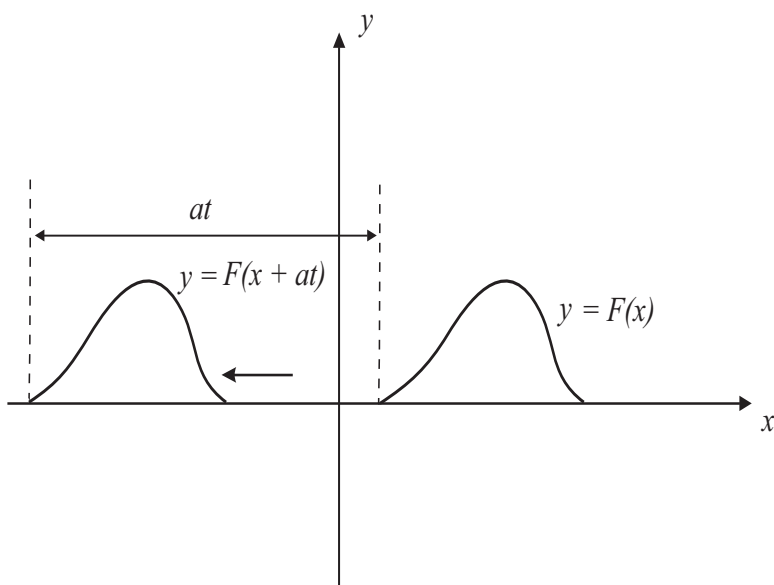
Figure 11.12: Wave of fixed shape moving to the left.

Exercises

1. Find the eigenvalues λ_n and the eigenfunctions y_n for the equation $y'' + \lambda y = 0$ in each of the following instances.

   (a) $y(0) = 0$, $y(\pi/2) = 0$

   (b) $y(0) = 0$, $y(2\pi) = 0$

   (c) $y(0) = 0$, $y(1) = 0$

   (d) $y(0) = 0$, $y(L) = 0$ for $L > 0$

   (e) $y(-L) = 0$, $y(L) = 0$ for $L > 0$

   (f) $y(a) = 0$, $y(b) = 0$ for $a < b$

   Solve the following two exercises without worrying about convergence of series or differentiability of functions.

2. If $y = F(x)$ is an arbitrary function, then $y = F(x + at)$ represents a wave of fixed shape that moves to the left along the x-axis with velocity a (Figure 11.12).

   Similarly, if $y = G(x)$ is another arbitrary function, then $y = G(x - at)$ is a wave moving to the right, and the most general one-dimensional wave with velocity a is

   $$y(x,t) = F(x + at) + G(x - at). \qquad (*)$$

   (a) Show that $(*)$ satisfies the wave equation.

   (b) It is easy to see that the constant a in the wave equation has the dimension of velocity. Also, it is intuitively clear that if a stretched string is disturbed,

then the waves will move in both directions away from the source of the disturbance. These considerations suggest introducing the new variables $\alpha = x + at$, $\beta = x - at$. Show that with these independent variables, the wave equation becomes

$$\frac{\partial^2 y}{\partial \alpha \partial \beta} = 0 \,.$$

From this derive $(*)$ by integration. Formula $(*)$ is called *d'Alembert's solution* of the wave equation. It was also obtained, slightly later and independently, by Euler.

3. Solve the vibrating string problem in the text if the initial shape $y(x, 0) = f(x)$ is specified by the given function. In each case, sketch the initial shape of the string on a set of axes.

(a)

$$f(x) = \begin{cases} 2cx/\pi & \text{if} & 0 \le x \le \pi/2 \\ 2c(\pi - x)/\pi & \text{if} & \pi/2 \le x \le \pi \end{cases}$$

(b)

$$f(x) = \frac{1}{\pi} x(\pi - x)$$

(c)

$$f(x) = \begin{cases} x & \text{if} & 0 \le x \le \pi/4 \\ \pi/4 & \text{if} & \pi/4 < x < 3\pi/4 \\ \pi - x & \text{if} & 3\pi/4 \le x \le \pi \end{cases}$$

4. Solve the vibrating string problem in the text if the initial shape $y(x, 0) = f(x)$ is that of a single arch of the sine curve $f(x) = c \sin x$. Show that the moving string always has the same general shape, regardless of the value of c. Do the same for functions of the form $f(x) = c \sin nx$. Show in particular that there are $n - 1$ points between $x = 0$ and $x = \pi$ at which the string remains motionless; these points are called *nodes*, and these solutions are called *standing waves*. Draw sketches to illustrate the movement of the standing waves.

5. The problem of the *struck string* is that of solving the wave equation with the boundary conditions

$$y(0, t) = 0, \quad y(\pi, t) = 0$$

and the initial conditions

$$\frac{\partial y}{\partial t}\bigg|_{t=0} = g(x) \qquad \text{and} \qquad y(x, 0) = 0 \,.$$

(These initial conditions mean that the string is initially in the equilibrium position, and has an initial velocity $g(x)$ at the point x as a result of being struck.) By separating variables and proceeding formally, obtain the solution

$$y(x, t) = \sum_{j=1}^{\infty} c_j \sin jx \sin jat \,,$$

where

$$c_j = \frac{2}{\pi ja} \int_0^{\pi} g(x) \sin jx \, dx \,.$$

6. Consider an infinite string stretched taut on the x-axis from $-\infty$ to $+\infty$. Let the string be drawn aside into a curve $y = f(x)$ and released, and assume that its subsequent motion is described by the wave equation.

   (a) Use $(*)$ in Exercise **2** to show that the string's displacement is given by *d'Alembert's formula*

$$y(x,t) = \frac{1}{2}[f(x+at) + f(x-at)].\qquad(**)$$

   Hint: Remember the initial conditions.

   (b) Assume further that the string remains motionless at the points $x = 0$ and $x = \pi$ (such points are called *nodes*), so that $y(0,t) = y(\pi,t) = 0$, and use $(**)$ to show that f is an odd function that is periodic with period 2π (that is, $f(-x) = f(x)$ and $f(x+2\pi) = f(x)$).

   (c) Show that since f is odd and periodic with period 2π then f necessarily vanishes at 0 and π.

   (d) Show that Bernoulli's solution of the wave equation can be written in the form $(**)$. **Hint:** $2\sin nx \cos nat = \sin[n(x+at)] + \sin[n(x-at)]$.

11.5   The Heat Equation

Fourier's Point of View

   In [FOU], Fourier considered variants of the following basic question. Let there be given an insulated, homogeneous rod of length π with initial temperature at each $x \in [0,\pi]$ given by a function $f(x)$ (Figure 11.13). Assume that the endpoints are held at temperature 0, and that the temperature of each cross-section is constant. The problem is to describe the temperature $u(x,t)$ of the point x in the rod at time t. Fourier perceived the fundamental importance of this problem as follows:

> Primary causes are unknown to us; but are subject to simple and constant laws, which may be discovered by observation, the study of them being the object of natural philosophy.
>
> Heat, like gravity, penetrates every substance of the universe, its rays occupying all parts of space. The object of our work is to set forth the mathematical laws which this element obeys. The theory of heat will hereafter form one of the most important branches of general physics.
>
> $\cdots\;\;\cdots$
>
> I have deduced these laws from prolonged study and attentive comparison of the facts known up to this time; all these facts I have observed afresh in the course of several years with the most exact instruments that have hitherto been used.

   Let us now describe the manner in which Fourier solved his problem. First, it is required to write a differential equation which u satisfies. We shall derive such an equation using three physical principles:
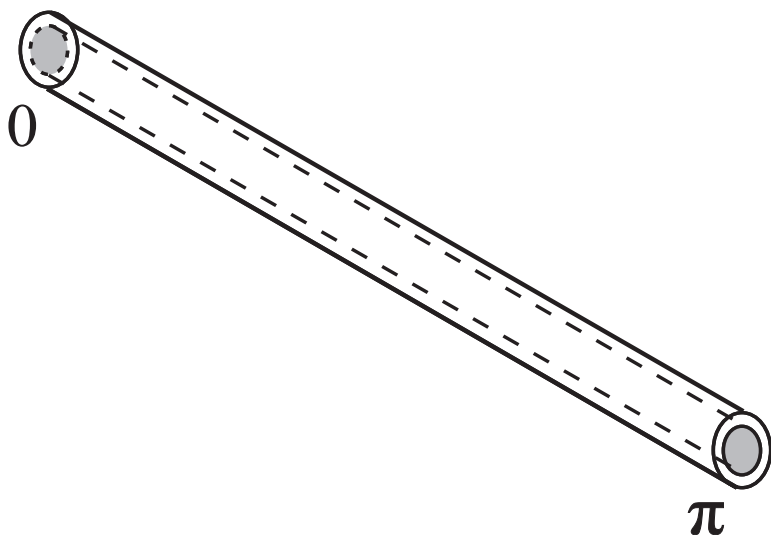
Figure 11.13: The insulated rod.

(1) The density of heat energy is proportional to the temperature u, hence the amount of heat energy in any interval $[a, b]$ of the rod is proportional to $\int_a^b u(x, t)\, dx$.

(2) (Newton's law of cooling) The rate at which heat flows from a hot place to a cold one is proportional to the difference in temperature. The infinitesimal version of this statement is that the rate of heat flow across a point x (from left to right) is some negative constant times $\partial_x u(x, t)$.

(3) (Conservation of Energy) Heat has no sources or sinks.

Now **(3)** tells us that the only way that heat can enter or leave any interval portion $[a, b]$ of the rod is through the endpoints. And **(2)** tells us exactly how this happens. Using **(1)**, we may therefore write

$$\frac{d}{dt} \int_a^b u(x, t)\, dx = \eta^2 [\partial_x u(b, t) - \partial_x u(a, t)].$$

We may rewrite this equation as

$$\int_a^b \partial_t u(x, t)\, dx = \eta^2 \int_a^b \partial_x^2 u(x, t)\, dx.$$

Differentiating in b, we find that

$$\partial_t u = \eta^2 \partial_x^2 u, \tag{11.5.1}$$

and that is the heat equation.

Suppose for simplicity that the constant of proportionality η^2 equals 1. Fourier guessed that equation (11.5.1) has a solution of the form $u(x,t) = \alpha(x)\beta(t)$. Substituting this guess into the equation yields

$$\alpha(x)\beta'(t) = \alpha''(x)\beta(t)$$

or

$$\frac{\beta'(t)}{\beta(t)} = \frac{\alpha''(x)}{\alpha(x)}.$$

Since the left side is independent of x and the right side is independent of t, it follows that there is a constant K such that

$$\frac{\beta'(t)}{\beta(t)} = K = \frac{\alpha''(x)}{\alpha(x)}$$

or

$$\begin{aligned} \beta'(t) &= K\beta(t) \\ \alpha''(x) &= K\alpha(x). \end{aligned}$$

We conclude that $\beta(t) = Ce^{Kt}$. The nature of β, and hence of α, thus depends on the sign of K. But physical considerations tell us that the temperature will dissipate as time goes on, so we conclude that $K \leq 0$. Therefore $\alpha(x) = \cos\sqrt{-K}\,x$ and $\alpha(x) = \sin\sqrt{-K}\,x$ are solutions of the differential equation for α. The initial conditions $u(0,t) = u(\pi,t) = 0$ (since the ends of the rod are held at constant temperature 0) eliminate the first of these solutions and force $K = -j^2$, j an integer. Thus Fourier found the solutions

$$u_j(x,t) = e^{-j^2 t}\sin jx, \quad j \in \mathbb{N}$$

of the heat equation. By linearity, any finite linear combination

$$u(x,t) = \sum_j b_j e^{-j^2 t}\sin jx \tag{11.5.2}$$

of these solutions is also a solution. It is plausible to extend this assertion to infinite linear combinations. Using the initial condition $u(x,0) = f(x)$ again raises the question of whether "any" function $f(x)$ on $[0,\pi]$ can be written as a (infinite) linear combination of the functions $\sin jx$.

Fourier's solution to this last problem (of the sine functions spanning essentially everything) is roughly as follows. Suppose f is a function that is so representable:

$$f(x) = \sum_j b_j \sin jx. \tag{11.5.3}$$

Setting $x = 0$ gives

$$f(0) = 0.$$

Differentiating both sides of (11.5.3) and setting $x = 0$ gives

$$f'(0) = \sum_{j=1}^{\infty} j b_j. \tag{11.5.4}$$

Successive differentiation of (11.5.3), and evaluation at 0, gives

$$f^{(k)}(0) = \sum_{j=1}^{\infty} j^k b_j (-1)^{\lfloor k/2 \rfloor}$$

for k odd (by oddness of f, the even derivatives must be 0 at 0). Here $\lfloor \quad \rfloor$ denotes the greatest integer function. Thus Fourier devised a system of infinitely many equations in the infinitely many unknowns $\{b_j\}$. He proceeded to solve this system by truncating it to an $N \times N$ system (the first N equations restricted to the first N unknowns), solving that truncated system, and then letting N tend to ∞. Suffice it to say that Fourier's arguments contained many dubious steps (see [FOU] and [LAN]).

The upshot of Fourier's intricate and lengthy calculations was that

$$b_j = \frac{2}{\pi} \int_0^{\pi} f(x) \sin jx \, dx. \tag{11.5.5}$$

By modern standards, Fourier's reasoning was specious; for he began by assuming that f possessed an expansion in terms of sine functions. The formula (11.5.5) hinges on that supposition, together with steps in which one compensated division by zero with a later division by ∞. Nonetheless, Fourier's methods give an actual *procedure* for endeavoring to expand any given f in a series of sine functions.

Fourier's abstract arguments constitute the first part of his book. The bulk, and remainder, of the book consists of separate chapters in which the expansions for particular functions are computed.

EXAMPLE 11.23 Suppose that the thin rod in the setup of the heat equation is first immersed in boiling water so that its temperature is uniformly 100°C. Then imagine that it is removed from the water at time $t = 0$ with its ends immediately put into ice so that these ends are kept at temperature 0°C. Let us find the temperature $u = u(x, t)$ under these circumstances.

The initial temperature distribution is given by the constant function

$$f(x) = 100 , \quad 0 < x < \pi .$$

The two boundary conditions, and the other initial condition, are as usual. Thus our job is simply this: to find the sine series expansion of this function f. We

calculate that

$$
\begin{aligned}
b_j &= \frac{2}{\pi} \int_0^\pi 100 \sin jx \, dx \\
&= -\frac{200}{\pi} \frac{\cos jx}{j} \Big|_0^\pi \\
&= -\frac{200}{\pi} \left[\frac{(-1)^j}{j} - \frac{1}{j} \right] \\
&= \begin{cases} 0 & \text{if} \quad j = 2\ell \text{ is even} \\[2mm] \dfrac{400}{\pi j} & \text{if} \quad j = 2\ell - 1 \text{ is odd}. \end{cases}
\end{aligned}
$$

Thus

$$
f(x) = \frac{400}{\pi} \left(\sin x + \frac{\sin 3x}{3} + \frac{\sin 5x}{5} + \cdots \right).
$$

Now, referring to formula (11.5.2) from our general discussion of the heat equation, we know that

$$
u(x, t) = \frac{400}{\pi} \left(e^{-t} \sin x + \frac{1}{3} e^{-9t} \sin 3x + \frac{1}{5} e^{-25t} \sin 5x + \cdots \right). \qquad \square
$$

EXAMPLE 11.24 Let us find the steady-state temperature of the thin rod from our analysis of the heat equation if the fixed temperatures at the ends $x = 0$ and $x = \pi$ are w_1 and w_2 respectively.

The phrase "steady state" means that $\partial u / \partial t = 0$, so that the heat equation reduced to $\partial^2 u / \partial x^2 = 0$ or $d^2 u / dx^2 = 0$. The general solution is then $u = Ax + B$. The values of these two constants A and B are forced by the two boundary conditions.

In fact a little high school algebra tells us that

$$
u = w_1 + \frac{1}{\pi} (w_2 - w_1) x. \qquad \square
$$

The steady-state version of the 3-dimensional heat equation

$$
a^2 \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} \right) = \frac{\partial u}{\partial t}
$$

is

$$
\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} = 0.
$$

This last is called *Laplace's equation*. The study of this equation and its solutions and subsolutions and their applications is a deep and rich branch of mathematics called *potential theory*. There are applications to heat, to gravitation, to electromagnetics, and to many other parts of physics. The equation plays a central role in the theory of partial differential equations, and is also an integral part of complex variable theory.

Exercises

1. Solve the boundary value problem

$$
\begin{aligned}
a^2\frac{\partial^2 w}{\partial x^2} &= \frac{\partial w}{\partial t} \\
w(x,0) &= f(x) \\
w(0,t) &= 0 \\
w(\pi,t) &= 0
\end{aligned}
$$

if the last three conditions—the boundary conditions—are changed to

$$
\begin{aligned}
w(x,0) &= f(x) \\
w(0,t) &= w_1 \\
w(\pi,t) &= w_2 \,.
\end{aligned}
$$

2. In the solution of the heat equation, suppose that the ends of the rod are insulated instead of being kept fixed at $0°C$. What are the new boundary conditions? Find the temperature $w(x,t)$ in this case by using just common sense.

3. Solve the problem of finding $w(x,t)$ for the rod with insulated ends at $x = 0$ and $x = \pi$ (see the preceding exercise) if the initial temperature distribution is given by $w(x,0) = f(x)$.

Solve the following two exercises without worrying about convergence of series or differentiability of functions.

4. Solve the Dirichlet problem for the unit disc when the boundary function $f(\theta)$ is defined by

(a) $f(\theta) = \cos\theta/2$, $\quad -\pi \le \theta \le \pi$

(b) $f(\theta) = \theta$, $\quad -\pi < \theta < 0$

(c) $f(\theta) = \begin{cases} 0 & \text{if} \quad -\pi \le \theta < 0 \\ \sin\theta & \text{if} \quad 0 \le \theta \le \pi \end{cases}$

(d) $f(\theta) = \begin{cases} 0 & \text{if} \quad -\pi \le \theta < \pi/2 \\ 1 & \text{if} \quad \pi/2 \le \theta \le \pi \end{cases}$

(e) $f(\theta) = \theta^2/4$, $\quad -\pi \le \theta \le \pi$

* **5.** Suppose that the lateral surface of the thin rod that we analyzed in the text is not insulated, but in fact radiates heat into the surrounding air. If Newton's law of cooling (that a body cools at a rate proportional to the difference of its temperature with the temperature of the surrounding

air) is assumed to apply, then show that the 1-dimensional heat equation becomes

$$a^2 \frac{\partial^2 w}{\partial x^2} = \frac{\partial w}{\partial t} + c(w - w_0)$$

where c is a positive constant and w_0 is the temperature of the surrounding air.

6. In Exercise 5, find $w(x, t)$ if the ends of the rod are kept at $0°C$, $w_0 = 0°C$, and the initial temperature distribution on the rod is $f(x)$.

* 7. The 2-dimensional heat equation is

$$a^2 \left(\frac{\partial^2 w}{\partial x^2} + \frac{\partial^2 w}{\partial y^2} \right) = \frac{\partial w}{\partial t}.$$

Use the method of separation of variables to find a steady-state solution of this equation in the infinite strip of the x-y plane bounded by the lines $x = 0$, $x = \pi$, and $y = 0$ if the following boundary conditions are satisfied:

$$w(0, y) = 0 \qquad\qquad\qquad w(\pi, y) = 0$$
$$w(x, 0) = f(x) \qquad\qquad\qquad \lim_{y \to +\infty} w(x, y) = 0.$$

* 8. Show that the Dirichlet problem for the disc $\{(x, y) : x^2 + y^2 \leq R^2\}$, where $f(\theta)$ is the boundary function, has the solution

$$w(r, \theta) = \frac{1}{2}a_0 + \sum_{j=1}^{\infty} \left(\frac{r}{R} \right)^j (a_j \cos j\theta + b_j \sin j\theta)$$

where a_j and b_j are the Fourier coefficients of f. Show also that the Poisson integral formula for this more general disc setting is

$$w(r, \theta) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{R^2 - r^2}{R^2 - 2Rr \cos(\theta - \phi) + r^2} f(\phi) \, d\phi.$$

* 9. Let w be a harmonic function in a planar region (that is, a function annihilated by the Laplacian), and let C be any circle entirely contained (along with its interior) in this region. Prove that the value of w at the center of C is the average of its values on the circumference.

Chapter 12

Functions of Several Variables

12.1   A New Look at the Basic Concepts of Analysis

A point of $\mathbb{R}^k$ is denoted $(x_1, x_2, \ldots, x_k)$. In the analysis of functions of one real variable, the domain of a function is typically an open interval. Since any open set in $\mathbb{R}^1$ is the disjoint union of open intervals, it is natural to work in the context of intervals. Such a simple situation does not obtain in the analysis of several variables. We will need some new notation and concepts in order to study functions in $\mathbb{R}^k$.

If $\mathbf{x} = (x_1, x_2, \ldots, x_k)$ is an element of $\mathbb{R}^k$, then we set

$$\|\mathbf{x}\| = \sqrt{(x_1)^2 + (x_2)^2 + \cdots + (x_k)^2}\,.$$

The expression $\|\mathbf{x}\|$ is commonly called the *norm* of $\mathbf{x}$. The norm of $\mathbf{x}$ measures the distance of $\mathbf{x}$ to the origin.

In general, we measure distance between two points $\mathbf{s} = (s_1, s_2, \ldots, s_k)$ and $\mathbf{t} = (t_1, t_2, \ldots, t_k)$ in $\mathbb{R}^k$ by the formula

$$\|\mathbf{s} - \mathbf{t}\| = \sqrt{(s_1 - t_1)^2 + (s_2 - t_2)^2 + \cdots + (s_k - t_k)^2}\,.$$

See Figure 12.1. Of course this notion of distance can be justified by considerations using the Pythagorean theorem (see the exercises), but we treat this as a definition. The distance between two points is nonnegative, and equals zero if and only if the two points are identical. Moreover, there is a triangle inequality:

$$\|\mathbf{s} - \mathbf{t}\| \le \|\mathbf{s} - \mathbf{u}\| + \|\mathbf{u} - \mathbf{t}\|\,.$$

We sketch a proof of this inequality in the exercises (by reducing it to the one-dimensional triangle inequality).
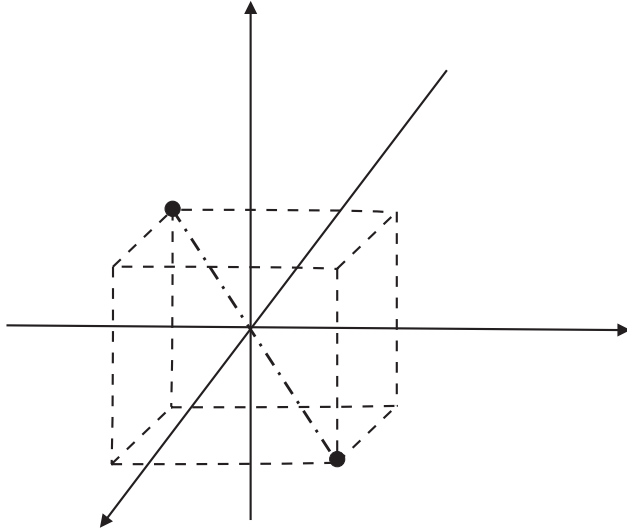
Figure 12.1: Distance in space.

Definition 12.1 If $x \in \mathbb{R}^k$ and $r > 0$, then the *open ball* with center x and radius r is the set

$$B(x, r) = \{t \in \mathbb{R}^k : \|x - t\| < r\}.$$

The *closed ball* with center x and radius r is the set

$$\overline{B}(x, r) = \{t \in \mathbb{R}^k : \|t - x\| \leq r\}.$$

Definition 12.2 A set $U \subseteq \mathbb{R}^k$ is said to be *open* if, for each $x \in U$, there is an $r > 0$ such that the ball $B(x, r)$ is contained in U.

EXAMPLE 12.3 Let

$$S = \{x = (x_1, x_2, x_3) \in \mathbb{R}^3 : 1 < \|x\| < 2\}.$$

This set is open. See Figure 12.2. For, if $x \in S$, let $r = \min\{\|x\| - 1, 2 - \|x\|\}$. Then $B(x, r)$ is contained in S for the following reason: if $t \in B(x, r)$ then

$$\|x\| \leq \|t - x\| + \|t\|$$

hence

$$\|t\| \geq \|x\| - \|t - x\| > \|x\| - r \geq \|x\| - (\|x\| - 1) = 1.$$

Likewise,

$$\|t\| \leq \|x\| + \|t - x\| < \|x\| + r \leq \|x\| + (2 - \|x\|) = 2.$$

It follows that $t \in S$ hence $B(x, r) \subseteq S$. We conclude that S is open.

However, a moment's thought shows that S could not be written as a disjoint union of open balls, or open cubes, or any other regular type of open set. □
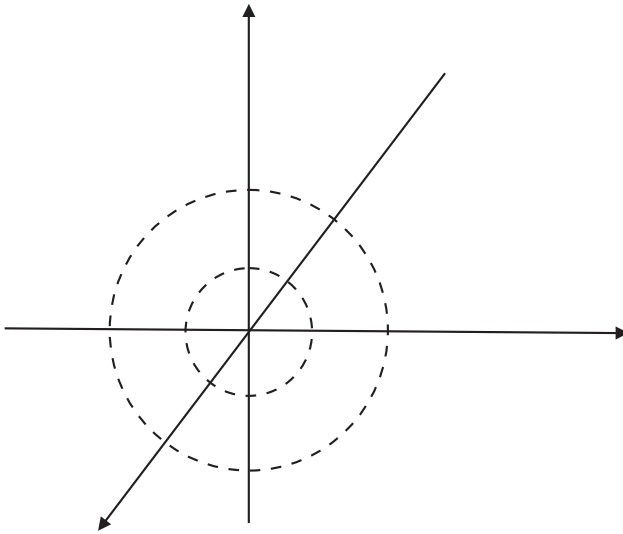
Figure 12.2: An open set.

In this chapter we consider functions with domain a set (usually open) in $\mathbb{R}^k$. See Figure 12.3. This means that the function f may be written in the form $f(x_1, x_2, \ldots, x_k)$. An example of such a function is $f(x_1, x_2, x_3, x_4) = x_1 \cdot (x_2)^4 - x_3/x_4$ or $g(x_1, x_2, x_3) = (x_3)^2 \cdot \sin(x_1 \cdot x_2 \cdot x_3)$.

Definition 12.4 Let $E \subseteq \mathbb{R}^k$ be a set and let f be a real-valued function with domain E. Fix a point $\mathbf{P}$ which is either in E or is an accumulation point of E (in the sense discussed in Chapter 4). We say that

$$\lim_{\boldsymbol{x} \to \mathbf{P}} f(\boldsymbol{x}) = \ell \,,$$

with ℓ a real number if, for each $\epsilon > 0$ there is a $\delta > 0$ such that, when $\boldsymbol{x} \in E$ and $0 < \|\boldsymbol{x} - \mathbf{P}\| < \delta$, then

$$|f(\boldsymbol{x}) - \ell| < \epsilon.$$

Refer to Figure 12.4.

Compare this definition with the definition in Section 5.1: the only difference is that we now measure the distance between points of the domain of f using $\| \ \|$ instead of $| \ |$.

EXAMPLE 12.5 The function

$$f(x_1, x_2, x_3) = \begin{cases} \dfrac{x_1 x_2}{x_1^2 + x_2^2 + x_3^2} & \text{if} \quad (x_1, x_2, x_3) \neq \mathbf{0} \\ 0 & \text{if} \quad (x_1, x_2, x_3) = \mathbf{0} \end{cases}$$
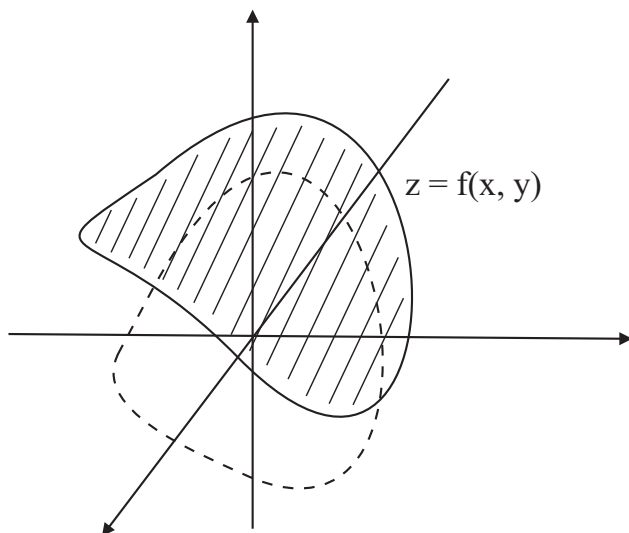
Figure 12.3: A function in space.
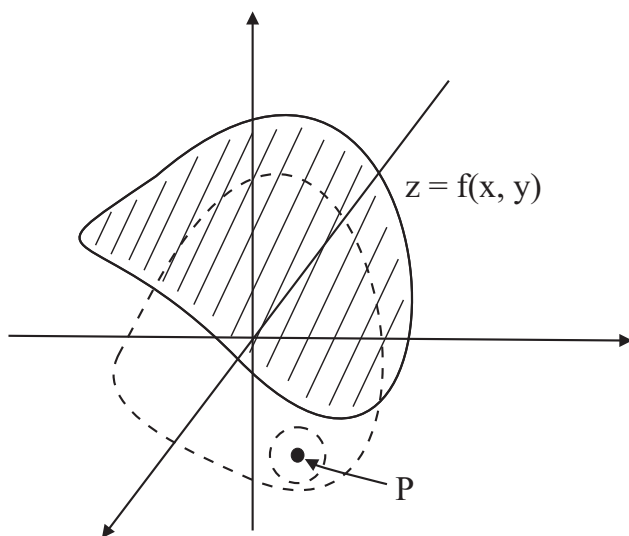
Figure 12.4: The limit of a function in space.
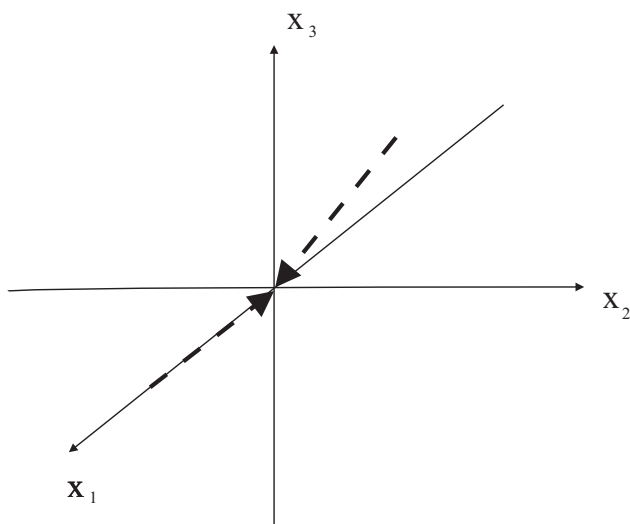
Figure 12.5: Approaching the origin from two different directions.

has no limit as $x \to 0$. For if we take $x = (t, 0, 0)$ then we obtain the limit

$$\lim_{t \to 0} f(t, 0, 0) = 0$$

while if we take $x = (t, t, t)$ then we obtain the limit

$$\lim_{t \to 0} f(t, t, t) = \frac{1}{3}.$$

Thus, for $\epsilon < \frac{1}{6} = \frac{1}{2} \cdot \frac{1}{3}$, there will exist no δ satisfying the definition of limit. See Figure 12.5.

However, the function

$$g(x_1, x_2, x_3, x_4) = x_1^2 + x_2^2 + x_3^2 + x_4^2$$

satisfies

$$\lim_{x \to 0} g(x) = 0$$

because, given $\epsilon > 0$, we take $\delta = \sqrt{\epsilon/4}$. Then $\|x - 0\| < \delta$ implies that $|x_j - 0| < \sqrt{\epsilon/4}$ for $j = 1, 2, 3, 4$ hence

$$|g(x_1, x_2, x_3, x_4) - 0| < \left| \left(\frac{\sqrt{\epsilon}}{\sqrt{4}} \right)^2 + \left(\frac{\sqrt{\epsilon}}{\sqrt{4}} \right)^2 + \left(\frac{\sqrt{\epsilon}}{\sqrt{4}} \right)^2 + \left(\frac{\sqrt{\epsilon}}{\sqrt{4}} \right)^2 \right| = \epsilon. \quad \square$$

Remark 12.6 Notice that, just as in the theory of one variable, the limit properties of f at a point P are independent of the *actual value* of f at P.
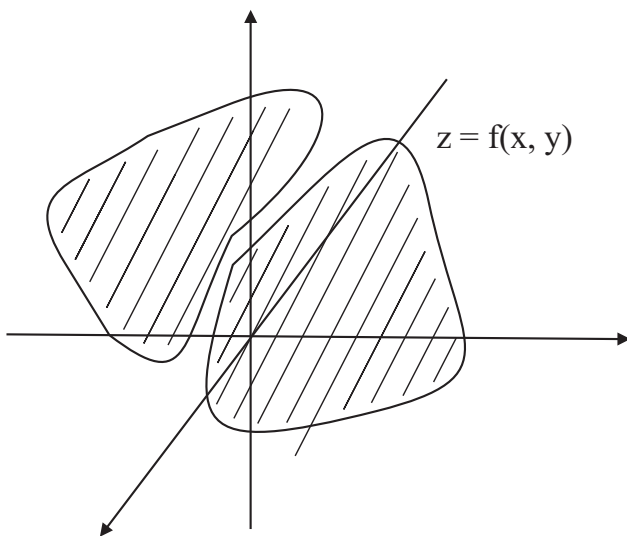
Figure 12.6: A discontinuous function.

Definition 12.7 Let f be a function with domain $E \subseteq \mathbb{R}^k$ and let $\mathbf{P} \in E$. We say that f is *continuous* at $\mathbf{P}$ if

$$\lim_{\boldsymbol{x} \to \mathbf{P}} f(\boldsymbol{x}) = f(\mathbf{P})\,.$$

See Figure 12.6.

The limiting process respects the elementary arithmetic operations, just as in the one-variable situation explored in Chapter 5. We will treat these matters in the exercises. Similarly, continuous functions are closed under the arithmetic operations (provided that we do not divide by zero). Next we turn to the fundamental properties of the derivative. [We refer the reader to Appendix III for a review of linear algebra.] In what follows, we use the notation ${}^t M$ to denote the transpose of the matrix M. We need the transpose so that the indicated matrix multiplications make sense.

Definition 12.8 Let $f(\boldsymbol{x})$ be a scalar-valued function whose domain contains a ball $B(\mathbf{P}, r)$. We say that f is *differentiable* at $\mathbf{P}$ if there is a $1 \times k$ matrix $M_{\mathbf{P}} = M_{\mathbf{P}}(f)$ such that, for all $\mathbf{h} \in \mathbb{R}^k$ satisfying $\|\mathbf{h}\| < r$, it holds that

$$f(\mathbf{P} + \mathbf{h}) = f(\mathbf{P}) + M_{\mathbf{P}} \cdot {}^t\mathbf{h} + \mathcal{R}_{\mathbf{P}}(f, \mathbf{h})\,, \qquad (12.8.1)$$

where

$$\lim_{\mathbf{h} \to \mathbf{0}} \frac{\mathcal{R}_{\mathbf{P}}(f, \mathbf{h})}{\|h\|} = 0\,.$$

The matrix $M_{\mathbf{P}} = M_{\mathbf{P}}(f)$ is called the *derivative* of f at $\mathbf{P}$.
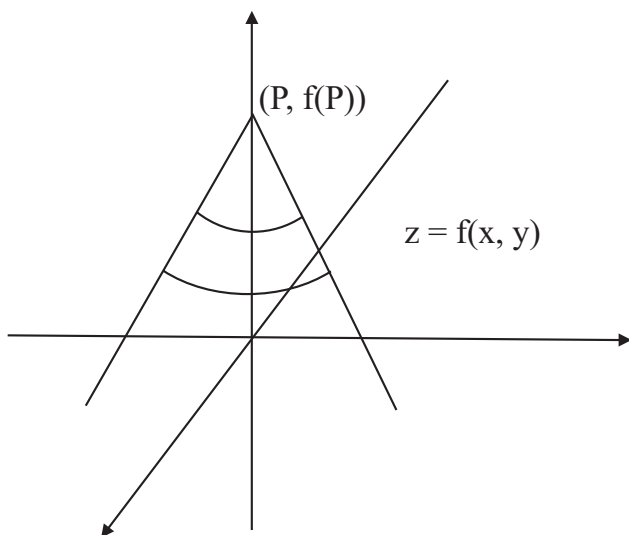
Figure 12.7: A function that is not differentiable at P.

EXAMPLE 12.9 Consider the scalar-valued function $f(x, y) = x^2 - 2xy$ at the point $\mathbf{P} = (1, 2)$. Let $\mathbf{h} = (h_1, h_2)$. The correct 1×2 matrix $M_{\mathbf{P}}$ is $(-2, -2)$ as we are about to see. This is because

$$
\begin{aligned}
f(\mathbf{P} + \mathbf{h}) &= f(P_1 + h_1, P_2 + h_2) \\
&= f(1 + h_1, 2 + h_2) \\
&= (1 + h_1)^2 - 2(1 + h_1)(2 + h_2) \\
&= [-3] + [-2h_1 - 2h_2] + [h_1^2 - 2h_1 h_2] \\
&= f(P) + M_{\mathbf{P}} \cdot {}^t\mathbf{h} + \mathcal{R}_{\mathbf{P}}(f, \mathbf{h}) \, .
\end{aligned}
$$

So we have verified that $M_{\mathbf{P}} = (-2, -2)$ is the derivative of f at $\mathbf{P}$. ☐

EXAMPLE 12.10 Consider the function $f(x, y) = 4 - \sqrt{x^2 + y^2}$. The graph of this function is the lower nappe of a cone. See Figure 12.7. It is easy to calculate, using $\mathbf{h} = (t, 0, 0)$ for $t < 0$ and $t > 0$, that this f is not differentiable at the origin. ☐

The best way to begin to understand any new idea is to reduce it to a situation that we already understand. If f is a function of one variable that is differentiable at $\mathbf{P} \in R$ then there is a number M such that

$$
\lim_{h \to 0} \frac{f(\mathbf{P} + h) - f(\mathbf{P})}{h} = M \, .
$$

We may rearrange this equality as

$$
\frac{f(\mathbf{P} + h) - f(\mathbf{P})}{h} - M = \mathcal{S}_{\mathbf{P}} ,
$$

where $\mathcal{S}_\mathbf{P} \to 0$ as $h \to 0$. But this may be rewritten as

$$f(\mathbf{P} + h) = f(\mathbf{P}) + M \cdot h + \mathcal{R}_\mathbf{P}(f, h), \tag{12.11}$$

where $\mathcal{R}_P = h \cdot \mathcal{S}_P$ and

$$\lim_{h \to 0} \frac{\mathcal{R}_\mathbf{P}(f, h)}{h} = 0 \,.$$

Equation (12.11) is parallel to (12.8.1) that defines the concept of derivative. The role of the $1 \times k$ matrix $M_\mathbf{P}$ is played here by the numerical constant M. *But a numerical constant is a 1×1 matrix.* Thus our equation in one variable is a special case of the equation in k variables. In one variable, the matrix representing the derivative is just the singleton consisting of the numerical derivative.

Note in passing that (just as in the one-variable case) the way that we now define the derivative of a function of several variables is closely related to the Taylor expansion. The number M in the one-variable case is the coefficient of the first order term in that expansion, which we know from Chapter 9 to be the first derivative.

What is the significance of the matrix $M_\mathbf{P}$ in our definition of derivative for a function of k real variables? Suppose that f is differentiable according to Definition 12.8. Let us attempt to calculate the "partial derivative" (as in calculus) with respect to x_1 of f. Let $\mathbf{h} = (h, 0, \ldots, 0)$. Then

$$f(P_1 + h, P_2, \ldots, P_k) = f(\mathbf{P}) + M_\mathbf{P} \cdot \begin{pmatrix} h \\ 0 \\ \vdots \\ 0 \end{pmatrix} + \mathcal{R}_\mathbf{P}(f, \mathbf{h}) \,.$$

Rearranging this equation we have

$$\frac{f(P_1 + h, P_2, \ldots, P_k) - f(\mathbf{P})}{h} = (M_\mathbf{P})_1 + \mathcal{S}_\mathbf{P} \,,$$

where $\mathcal{S}_\mathbf{P} \to 0$ as $h \to 0$ and $(M_\mathbf{P})_1$ is the first entry of the $1 \times k$ matrix $M_\mathbf{P}$.

But, letting $h \to 0$ in this last equation, we see that the partial derivative with respect to x_1 of the function f exists at P and equals $(M_\mathbf{P})_1$. A similar calculation shows that the partial derivative with respect to x_2 of the function f exists at P and equals $(M_\mathbf{P})_2$; likewise the partial derivative with respect to x_j of the function f exists at P and equals $(M_\mathbf{P})_j$ for $j = 1, \ldots, k$.

We summarize with a theorem:

Theorem 12.11 *Let f be a function defined on an open ball $B(\mathbf{P}, r) \subseteq R^k$ and suppose that f is differentiable at $\mathbf{P}$ with derivative the $1 \times k$ matrix $M_\mathbf{P}$. Then the first partial derivatives of f at $\mathbf{P}$ exist and they are, respectively, the entries of $M_\mathbf{P}$. That is,*

$$(M_\mathbf{P})_1 = \frac{\partial}{\partial x_1} f(\mathbf{P}) \;, \quad (M_\mathbf{P})_2 = \frac{\partial}{\partial x_2} f(\mathbf{P}) \;, \quad \ldots \;, \quad (M_\mathbf{P})_k = \frac{\partial}{\partial x_k} f(\mathbf{P}) \,.$$

EXAMPLE 12.12 Let $f(x, y) = \sin x - 3y$ and let $P = (0, 0)$. Then

$$\frac{\partial f}{\partial x}(P) = \cos x \Big|_{x=0, y=0} = 1$$

and

$$\frac{\partial f}{\partial y}(P) = -3\,.$$

So we see that $M_{\mathbf{P}} = (1, -3)$ and we are guaranteed that

$$f(\mathbf{P} + \mathbf{h}) = f(\mathbf{P}) + M_{\mathbf{P}} \cdot {}^t\mathbf{h} + \mathcal{R}_{\mathbf{P}}(f, \mathbf{h})\,. \qquad \Box$$

Unfortunately the converse of this theorem is not true: it is possible for the first partial derivatives of f to exist at a single point $\mathbf{P}$ without f being differentiable at $\mathbf{P}$ in the sense of Definition 12.8. Counterexamples will be explored in the exercises. On the other hand, as the last example suggests, the two different notions of *continuous differentiablity* are the same. We formalize this statement with a proposition:

Proposition 12.13 *Let f be a function defined on an open ball $B(\mathbf{P}, r)$. Assume that f is differentiable at each point of $B(\mathbf{P}, r)$ in the sense of Definition 12.8 and that the function*

$$\boldsymbol{x} \mapsto M_{\boldsymbol{x}}$$

is continuous in the sense that each of the functions

$$\boldsymbol{x} \mapsto (M_{\boldsymbol{x}})_j$$

is continuous, $j = 1, 2, \ldots, k$. Then each of the partial derivatives

$$\frac{\partial}{\partial x_1} f(\boldsymbol{x}) \quad \frac{\partial}{\partial x_2} f(\boldsymbol{x}) \quad \cdots \quad \frac{\partial}{\partial x_k} f(\boldsymbol{x})$$

exists for $\boldsymbol{x} \in B(\mathbf{P}, r)$ and is continuous.

Conversely, if each of the partial derivatives exists on $B(\mathbf{P}, r)$ and is continuous at each point then $M_{\boldsymbol{x}}$ exists at each point $\boldsymbol{x} \in B(\mathbf{P}, r)$ and is continuous. The entries of M_x are given by the partial derivatives of f.

Proof: This is essentially a routine check of definitions. The only place where the continuity is used is in proving the converse: that the existence and continuity of the partial derivatives implies the existence of $M_{\boldsymbol{x}}$. In proving the converse you should apply the one-variable Taylor expansion to the function $t \mapsto f(x + t\mathbf{h})$. $\qquad \Box$

Exercises

1. Fix elements $\mathbf{s}, \mathbf{t}, \mathbf{u} \in \mathbb{R}^k$. First assume that these three points are colinear. By reduction to the one-dimensional case, prove the Triangle Inequality

$$\|\mathbf{s} - \mathbf{t}\| \leq \|\mathbf{s} - \mathbf{u}\| + \|\mathbf{u} - \mathbf{t}\|.$$

   Now establish the general case of the Triangle Inequality by comparison with the colinear case.

2. Give another proof of the triangle inequality by squaring both sides and invoking the Schwarz inequality.

3. Formulate and prove the elementary properties of limits for functions of k variables (refer to Chapter 5 for the one-variable analogues).

4. Give an example of an infinitely differentiable function with domain $\mathbb{R}^2$ such that $\{(x_1, x_2) : f(x_1, x_2) = 0\} = \{(x_1, x_2) : |x_1|^2 + |x_2|^2 \leq 1\}$.

5. Formulate a notion of uniform convergence for functions of k real variables. Prove that the uniform limit of a sequence of continuous functions is continuous.

6. Formulate a notion of "compact set" for subsets of $\mathbb{R}^k$. Prove that the continuous image, under a vector-valued function, of a compact set is compact.

7. Refer to Exercise **6**. Prove that if f is a continuous function on a compact set then f assumes both a maximum value and a minimum value.

8. Give an example of a connected set in $\mathbb{R}^2$ with disconnected boundary.

9. Give an example of a disconnected set in $\mathbb{R}^2$ with infinitely many connected components.

10. Justify our notion of distance in $\mathbb{R}^k$ using Pythagorean Theorem considerations.

11. If $\mathbf{s}, \mathbf{t} \in \mathbb{R}^k$ then prove that

$$\|\mathbf{s} + \mathbf{t}\| \geq \|\mathbf{s}\| - \|\mathbf{t}\|.$$

12. Prove Proposition 12.13.

* 13. Give an example of a function f of two variables such that f has both first partial derivatives at a point P, yet f is not differentiable at P according to Definition 12.8.

12.2 Properties of the Derivative

The arithmetic properties of the derivative—that is the sum and difference, scalar multiplication, product, and quotient rules—are straightforward and are left to the exercises for you to consider. However, the Chain Rule takes on a different form and requires careful consideration.

In order to treat meaningful instances of the Chain Rule, we must first discuss *vector-valued* functions. That is, we consider functions with domain a subset of $\mathbb{R}^k$ and range *either* $\mathbb{R}^1$ *or* $\mathbb{R}^2$ *or* $\mathbb{R}^m$ for some integer $m > 0$. When we consider vector-valued functions, it simplifies notation if we consider all vectors to be column vectors. This convention will be in effect for the rest of the chapter. (Thus we will no longer use the "transpose" notation.) Note in passing that the expression $\|\boldsymbol{x}\|$ means the same thing for a column vector as it does for a row vector—the square root of the sum of the squares of the components. Also $f(\boldsymbol{x})$ means the same thing whether $\boldsymbol{x}$ is written as a row vector or a column vector.

EXAMPLE 12.14 Define the function

$$f(x_1, x_2, x_3) = \left(\begin{array}{c} (x_1)^2 - x_2 \cdot x_3 \\ \\ x_1 \cdot (x_2)^3 \end{array} \right).$$

This is a function with domain consisting of all triples of real numbers, or $\mathbb{R}^3$, and range consisting of all pairs of real numbers, or $\mathbb{R}^2$. For example,

$$f(-1, 2, 4) = \left(\begin{array}{c} -7 \\ -8 \end{array} \right). \qquad \Box$$

We say that a vector-valued function of k variables

$$f(\boldsymbol{x}) = (f_1(\boldsymbol{x}), f_2(\boldsymbol{x}), \ldots, f_m(\boldsymbol{x}))$$

(where m is a positive integer) is differentiable at a point $\mathbf{P}$ if each of its component functions is differentiable in the sense of Section 1. For example, the function

$$f(x_1, x_2, x_3) = \left(\begin{array}{c} x_1 \cdot x_2 \\ (x_3)^2 \end{array} \right)$$

is differentiable at all points (because $f_1(x_1, x_2, x_3) = x_1 \cdot x_2$ and $f_2(x_1, x_2, x_3) = x_3^2$ are differentiable) while the function

$$g(x_1, x_2, x_3) = \left(\begin{array}{c} x_2 \\ |x_3| - x_1 \end{array} \right)$$

is not differentiable at points of the form $(x_1, x_2, 0)$.

It is a good exercise in matrix algebra (which you will be asked to do at the end of the section) to verify that a vector-valued function f is differentiable at a point $\mathbf{P}$ if and only if there is an $m \times k$ matrix (where k is the dimension of the domain and m the dimension of the range) $M_{\mathbf{P}}(f)$ such that

$$f(\mathbf{P} + \mathbf{h}) = f(\mathbf{P}) + M_{\mathbf{P}}(f)\mathbf{h} + \mathcal{R}_{\mathbf{P}}(f, \mathbf{h});$$

here $\mathbf{h}$ is a $\mathbf{k}$-column vector and the remainder term $\mathcal{R}_{\mathbf{P}}$ is a column vector satisfying

$$\frac{\|\mathcal{R}_{\mathbf{P}}(f, \mathbf{h})\|}{\|\mathbf{h}\|} \to 0$$

as $\mathbf{h} \to 0$. One nice consequence of this formula is that, by what we learned in the last section about partial derivatives, the entry in the ith row and j^{th} column of the matrix $M_{\mathbf{P}}(f)$ is $\partial f_i / \partial x_j$. Here f_i is the ith component of the mapping f.

Of course the Chain Rule provides a method for differentiating compositions of functions. What we will discover in this section is that the device of thinking of the derivative as a matrix occurring in an expansion of f about a point $\mathbf{P}$ makes the Chain Rule a very natural and easy result to derive. It will also prove to be a useful way of keeping track of information.

Theorem 12.15 *Let g be a function of k real variables taking values in $\mathbb{R}^m$ and let f be a function of m real variables taking values in $\mathbb{R}^n$. Suppose that the range of g is contained in the domain of f, so that $f \circ g$ makes sense. If g is differentiable at a point $\mathbf{P}$ in its domain and f is differentiable at $g(\mathbf{P})$ then $f \circ g$ is differentiable at $\mathbf{P}$ and its derivative is $M_{g(\mathbf{P})}(f) \cdot M_{\mathbf{P}}(g)$. We use the symbol $\cdot$ here to denote matrix multiplication.*

Proof: By the hypothesis about the differentiability of g,

$$\begin{aligned}
(f \circ g)(\mathbf{P} + \mathbf{h}) &= f(g(\mathbf{P} + \mathbf{h})) \\
&= f\left(g(\mathbf{P}) + M_{\mathbf{P}}(g)\mathbf{h} + \mathcal{R}_{\mathbf{P}}(g, \mathbf{h})\right) \\
&= f\left(g(\mathbf{P}) + \mathbf{k}\right), \quad\quad\quad (12.15.1)
\end{aligned}$$

where

$$\mathbf{k} = M_{\mathbf{P}}(g)\mathbf{h} + \mathcal{R}_{\mathbf{P}}(g, \mathbf{h}).$$

But then the differentiability of f at $g(\mathbf{P})$ implies that (12.15.1) equals

$$f(g(\mathbf{P})) + M_{g(\mathbf{P})}(f)\mathbf{k} + \mathcal{R}_{g(\mathbf{P})}(f, \mathbf{k}).$$

Now let us substitute in the value of $\mathbf{k}$. We find that

$$\begin{aligned}
(f \circ g)(\mathbf{P} + \mathbf{h}) &= f(g(\mathbf{P})) + M_{g(\mathbf{P})}(f)[M_{\mathbf{P}}(g)\mathbf{h} + \mathcal{R}_{\mathbf{P}}(g, \mathbf{h})] \\
&\quad\quad + \mathcal{R}_{g(\mathbf{P})}(f, M_{\mathbf{P}}(g)\mathbf{h} + \mathcal{R}_{\mathbf{P}}(g, \mathbf{h})) \\
&= f(g(\mathbf{P})) + M_{g(\mathbf{P})}(f)M_{\mathbf{P}}(g)\mathbf{h} \\
&\quad\quad + \big\{ M_{g(\mathbf{P})}(f)\mathcal{R}_{\mathbf{P}}(g, \mathbf{h}) \\
&\quad\quad + \mathcal{R}_{g(\mathbf{P})}(f, M_{\mathbf{P}}(g)\mathbf{h} + \mathcal{R}_{\mathbf{P}}(g, \mathbf{h})) \big\} \\
&\equiv f(g(\mathbf{P})) + M_{g(\mathbf{P})}(f)M_{\mathbf{P}}(g)\mathbf{h} \\
&\quad\quad + \mathcal{Q}_{\mathbf{P}}(f \circ g, \mathbf{h}),
\end{aligned}$$

where the last equality *defines* $\mathcal{Q}$. The term $\mathcal{Q}$ should be thought of as a remainder term. Since

$$\frac{\|\mathcal{R}_{\mathbf{P}}(g, \mathbf{h})\|}{\|\mathbf{h}\|} \to 0$$

as $\mathbf{h} \to 0$, it follows that

$$\frac{M_{g(\mathbf{P})}(f)\mathcal{R}_{\mathbf{P}}(g, \mathbf{h})}{\|\mathbf{h}\|} \to 0 \,.$$

(Details of this assertion are requested of you in the exercises.) Similarly,

$$\frac{\mathcal{R}_{g(\mathbf{P})}(f, M_{\mathbf{P}}(g)\mathbf{h} + \mathcal{R}_{\mathbf{P}}(g, \mathbf{h}))}{\|\mathbf{h}\|} \to 0$$

as $\mathbf{h} \to 0$.

In conclusion, we see that $f \circ g$ is differentiable at $\mathbf{P}$ and that the derivative equals $M_{g(\mathbf{P})}(f)M_{\mathbf{P}}(g)$, the product of the derivatives of f and g. $\qquad\square$

Remark 12.16 Notice that, by our hypotheses, $M_{\mathbf{P}}(g)$ is an $m \times k$ size matrix and $M_{g(\mathbf{P})}(f)$ is an $n \times m$ size matrix. Thus their product makes sense.

In general, if g is a function from a subset of $\mathbb{R}^k$ to $\mathbb{R}^m$ then, if we want $f \circ g$ to make sense, f must be a function from a subset of $\mathbb{R}^m$ to some $\mathbb{R}^n$. In other words, the dimension of the range of g had better match the dimension of the domain of f. Then the derivative of g at some point $\mathbf{P}$ will be an $m \times k$ matrix and the derivative of f at $g(\mathbf{P})$ will be an $n \times m$ matrix. Hence the matrix multiplication $M_{g(\mathbf{P})}(f)M_{\mathbf{P}}(g)$ will make sense.

Corollary 12.17 (The Chain Rule in Coordinates) *Let* $f : \mathbb{R}^m \to \mathbb{R}^n$ *and* $g : \mathbb{R}^k \to \mathbb{R}^m$ *be vector-valued functions and assume that* $h = f \circ g$ *makes sense. If* g *is differentiable at a point* $\mathbf{P}$ *of its domain and* f *is differentiable at* $g(\mathbf{P})$ *then for each* i *and* j *we have*

$$\frac{\partial h_i}{\partial x_j}(\mathbf{P}) = \sum_{\ell=1}^{m} \frac{\partial f_i}{\partial s_\ell}(g(\mathbf{P})) \cdot \frac{\partial g_\ell}{\partial x_j}(\mathbf{P}) \,.$$

Proof: The function $\partial h_i/x_j$ is the entry of $M_{\mathbf{P}}(h)$ in the ith row and j^{th} column. However, $M_{\mathbf{P}}(h)$ is the product of $M_{g(\mathbf{P})}(f)$ with $M_{\mathbf{P}}(g)$ by Theorem 12.15. The entry in the ith row and jth column of that product is

$$\sum_{\ell=1}^{m} \frac{\partial f_i}{\partial s_\ell}(g(\mathbf{P})) \cdot \frac{\partial g_\ell}{\partial x_j}(\mathbf{P}). \qquad\square$$

EXAMPLE 12.18 Let $f(x, y) = x^2 - y^2$ and let $g(s, t) = (st, -t^3)$. Then $f \circ g$ makes sense and we may calculate the derivative of this composition at the point $P = (1, 3)$. Let us first do so according to the matrix rule given in Theorem

12.15. And then let us follow that with the analogous calculation in coordinates (as in the corollary).

We write $g(s, t) = (st, -t^3) = (g_1(s, t), g_2(s, t))$. Now begin by calculating

$$\frac{\partial g_1}{\partial s} = t \qquad \text{and} \qquad \frac{\partial g_1}{\partial t} = s$$

and

$$\frac{\partial g_2}{\partial s} = 0 \qquad \text{and} \qquad \frac{\partial g_2}{\partial t} = -3t^3 \,.$$

Thus

$$\frac{\partial g_1}{\partial s}(\mathbf{P}) = 3 \qquad \text{and} \qquad \frac{\partial g_1}{\partial t}(\mathbf{P}) = 1$$

and

$$\frac{\partial g_2}{\partial s}(\mathbf{P}) = 0 \qquad \text{and} \qquad \frac{\partial g_2}{\partial t}(\mathbf{P}) = -27 \,.$$

Therefore

$$M_{\mathbf{P}}(g) = \begin{pmatrix} 3 & 1 \\ 0 & -27 \end{pmatrix} \,.$$

Next we note that $g(P) = (3, -27)$ and

$$\frac{\partial f}{\partial x} = 2x \qquad \text{and} \qquad \frac{\partial f}{\partial y} = -2y$$

so that

$$\frac{\partial f}{\partial x}(g(\mathbf{P})) = 6 \qquad \text{and} \qquad \frac{\partial f}{\partial y}(g(\mathbf{P})) = 54 \,.$$

In conclusion,

$$M_{\mathbf{P}}(f \circ g) = (6, 54) \cdot \begin{pmatrix} 3 & 1 \\ 0 & -27 \end{pmatrix} = (18, -1452) \,.$$

Now let us perform the same calculation in coordinates. We begin by writing

$$f \circ g(s, t) = (st)^2 - (-t^3)^2 = s^2 t^2 - t^6 \,.$$

Now we calculate that

$$\frac{\partial (f \circ g)}{\partial s} = \frac{\partial f}{\partial x} \cdot \frac{\partial g_1}{\partial s} + \frac{\partial f}{\partial y} \cdot \frac{\partial g_2}{\partial s} = 2x \cdot 3 + (-2y) \cdot 0 = 2 \cdot 3 \cdot 3 = 18$$

and

$$\frac{\partial (f \circ g)}{\partial t} = \frac{\partial f}{\partial x} \cdot \frac{\partial g_1}{\partial t} + \frac{\partial f}{\partial y} \cdot \frac{\partial g_2}{\partial t} = 2x \cdot 1 + (-2y) \cdot (-27) = 6 - 1458 = -1452 \,.$$

In sum, the two methods of calculation give the same answer.   □

We conclude this section by deriving a Taylor expansion for scalar-valued functions of k real variables: this expansion for functions of several variables is derived in an interesting way from the expansion for functions of one variable. We say that a function f of several real variables is k times continuously differentiable if all partial derivatives of orders up to and including k exist and are continuous on the domain of f.

Theorem 12.19 (Taylor's Expansion) *For q a nonnegative integer let f be a $q + 1$ times continuously differentiable scalar-valued function on a neighborhood of a closed ball $\overline{B}(\mathbf{P}, r) \subseteq \mathbb{R}^k$. Then, for $x \in B(\mathbf{P}, r)$,*

$$f(\boldsymbol{x})$$
$$= \sum_{0 \leq j_1 + j_2 + \cdots + j_k \leq q} \frac{\partial^{j_1 + j_2 + \cdots + j_k} f}{\partial x_1^{j_1} \partial x_2^{j_2} \cdots \partial x_k^{j_k}}(\mathbf{P}) \cdot \frac{(x_1 - P_1)^{j_1}(x_2 - P_2)^{j_2} \cdots (x_k - P_k)^{j_k}}{(j_1)!(j_2)! \cdots (j_k)!}$$

$$+ \quad \mathcal{R}_{q,\mathbf{P}}(\boldsymbol{x}),$$

where

$$|\mathcal{R}_{q,\mathbf{P}}(x)| \leq C_0 \cdot \frac{\|\boldsymbol{x} - \mathbf{P}\|^{q+1}}{(q+1)!},$$

and

$$C_0 = \sup_{\substack{\mathbf{s} \in \overline{B}(\mathbf{P}, r) \\ \ell_1 + \ell_2 + \cdots + \ell_k = q+1}} \left| \frac{\partial^{j_1 + j_2 + \cdots + j_k} f}{\partial x_1^{j_1} \partial x_2^{j_2} \cdots \partial x_k^{j_k}}(s) \right|.$$

Proof: With $\mathbf{P}$ and $\boldsymbol{x}$ fixed, define

$$\mathcal{F}(s) = f(\mathbf{P} + s(\boldsymbol{x} - \mathbf{P})), \quad 0 \leq s < \frac{r}{\|\boldsymbol{x} - \mathbf{P}\|}.$$

We apply the one-dimensional Taylor theorem to the function $\mathcal{F}$, expanded about the point 0:

$$\mathcal{F}(s) = \sum_{\ell=0}^{q} \mathcal{F}^{(\ell)}(0) \frac{s^\ell}{\ell!} + R_{q,0}(\mathcal{F}, s).$$

Now the Chain Rule shows that

$$\mathcal{F}^{(\ell)}(0) =$$
$$\sum_{j_1 + j_2 + \cdots + j_k = \ell} \frac{\partial^{j_1 + j_2 + \cdots + j_k} f}{\partial x_1^{j_1} \partial x_2^{j_2} \cdots \partial x_k^{j_k}}(\mathbf{P})$$
$$\cdot \frac{\ell!}{(j_1)!(j_2)! \cdots (j_k)!} \cdot (x_1 - P_1)^{j_1}(x_2 - P_2)^{j_2} \cdots (x_k - P_k)^{j_k}.$$

Substituting this last equation, for each ℓ, into the formula for $\mathcal{F}(s)$ and setting $s = 1$ (recall that $r/\|\boldsymbol{x} - \mathbf{P}\| > 1$ since $x \in B(\mathbf{P}, r)$) yields the desired expression for $f(x)$. It remains to estimate the remainder term.

The one-variable Taylor theorem tells us that, for $s > 0$,

$$
\begin{aligned}
|R_{q,0}(\mathcal{F}, s)| &= \left| \int_0^s \mathcal{F}^{(q+1)}(\sigma) \frac{(s - \sigma)^q}{q!} d\sigma \right| \\
&\leq \int_0^s C_0 \cdot \|x - \mathbf{P}\|^{q+1} \cdot \left| \frac{(s - \sigma)^q}{q!} \right| d\sigma \\
&= C_0 \cdot \frac{\|x - \mathbf{P}\|^{q+1}}{(q+1)!} \, .
\end{aligned}
$$

Here we have of course used the Chain Rule to pass from derivatives of $\mathcal{F}$ to derivatives of f. This is the desired result. $\qquad\square$

EXAMPLE 12.20 Let us determine the degree-three Taylor expansion for the function $f(x, y) = x \cos y$ expanded about the point $P = (0, 0)$.

Following the theorem, we calculate as follows:

$$f(P) = 0 \, ,$$

$$\frac{\partial f}{\partial x}(P) = \cos y \Big|_{(0,0)} = 1 \, ,$$

$$\frac{\partial f}{\partial y}(P) = -x \sin y \Big|_{(0,0)} = 0 \, ,$$

$$\frac{\partial^2 f}{\partial x^2}(P) = 0 \, ,$$

$$\frac{\partial^2 f}{\partial y^2}(P) = -x \cos y \Big|_{(0,0)} = 0 \, ,$$

$$\frac{\partial^2 f}{\partial x \partial y}(P) = -\sin y \Big|_{(0,0)} = 0 \, ,$$

$$\frac{\partial^3 f}{\partial x^3}(P) = 0 \, ,$$

$$\frac{\partial^3 f}{\partial x^2 \partial y}(P) = 0 \, ,$$

$$\frac{\partial^3 f}{\partial x \partial y^2} = -\cos y \Big|_{(0,0)} = -1 \, ,$$

$$\frac{\partial^3 f}{\partial y^3} = x \sin y \Big|_{(0,0)} = 0 \, .$$

We find, then, that the Taylor expansion is

$$
\begin{aligned}
f(x, y) \;=\;\; & 0 + 1(x - 0) + 0(y - 0) \\
& + \frac{1}{2!}\left(0(x-0)^2 + 2\cdot 0(x-0)(y-0) + 0(y-0)^2\right) \\
& + \frac{1}{3!}\left(0(x-0)^3 + 3\cdot 0(x-0)^2(y-0) + 3\cdot(-1)(x-0)(y-0)^2\right. \\
& \left. + 0(y-0)^3\right) + \mathcal{R}_{3,\mathbf{P}}(\mathbf{x})\,.
\end{aligned}
$$

This of course simplifies to

$$
f(x,y) = x - \frac{3}{3!}xy^2 + \mathcal{R}_{3,\mathbf{P}}(\mathbf{x})\,. \qquad\qquad \square
$$

Exercises

1. Formulate a sum, product, and quotient rule for functions of two real variables taking values in $\mathbb{R}$.

2. Use the Chain Rule for a function $f : \mathbb{R}^n \to \mathbb{R}^n$ to find a formula for the derivative of the inverse of f in terms of the derivative of f itself.

3. Formulate a definition of second derivative parallel to the definition of first derivative given in Section 12.1. Your definition should involve a matrix. What does this matrix tell us about the second partial derivatives of the function?

4. If f and g are vector-valued functions with domain $\mathbb{R}^k$, both taking values in $\mathbb{R}^m$ and both having the same domain, then we can define the dot product function $h(\boldsymbol{x}) = f(\boldsymbol{x}) \cdot g(\boldsymbol{x})$. Formulate and prove a derivative Product Rule for this type of product.

5. Prove that if a function with domain an open subset of $\mathbb{R}^k$ is differentiable at a point P then it is continuous at P.

6. Let f be a function defined on a ball $B(\mathbf{P}, r)$. Let $\mathbf{u} = (u_1, u_2, \ldots, u_k)$ be a vector of unit length. If f is differentiable at $\mathbf{P}$ then give a definition of the directional derivative $D_{\mathbf{u}}f(\mathbf{P})$ of f in the direction $\mathbf{u}$ at P in terms of $M_{\mathbf{P}}$.

7. If f is differentiable on a ball $B(\mathbf{P}, r)$ and if $M_{\boldsymbol{x}}$ is the zero matrix for every $\boldsymbol{x} \in B(\mathbf{P}, r)$ then prove that f is constant on $B(\mathbf{P}, r)$.

8. Refer to Exercise **6** for notation. For which collections of vectors $\mathbf{u}_1, \mathbf{u}_2, \ldots,$ $\mathbf{u}_k$ in $\mathbb{R}^k$ is it true that if $D_{\mathbf{u}_j}f(x) = 0$ for all $x \in B(\mathbf{P}, r)$ and all $j = 1, 2, \ldots, k$ then f is identically constant?

9. Prove that an $\mathbb{R}^m$-valued function $\mathbf{f}$ is differentiable at a point $\mathbf{P} \in \mathbb{R}^k$ if and only if there is an $m \times k$ matrix (where k is the dimension of the domain and m the dimension of the range) $M_{\mathbf{P}}(\mathbf{f})$ such that

$$\mathbf{f}(\mathbf{P} + \mathbf{h}) = \mathbf{f}(\mathbf{P}) + M_{\mathbf{P}}(\mathbf{f})\mathbf{h} + \mathcal{R}_{\mathbf{P}}(\mathbf{f}, \mathbf{h}) \, ;$$

here $\mathbf{h}$ is a k-column vector and the remainder term $\mathcal{R}_{\mathbf{P}}$ is a column vector satisfying

$$\frac{\|\mathcal{R}_{\mathbf{P}}(\mathbf{f}, \mathbf{h})\|}{\|\mathbf{h}\|} \to 0$$

as $\mathbf{h} \to 0$.

10. Verify the last assertion in the proof of Theorem 12.15.

11. Prove, in the context of two real variables, that the composition of two continuously differentiable mappings is continuously differentiable.

12. Prove that the product of continuously differentiable functions is continuously differentiable.

* 13. There is no mean value theorem as such in the theory of functions of several real variables. For example, if $\gamma : [0, 1] \to \mathbb{R}^k$ is a differentiable function on $(0, 1)$, continuous on $[0, 1]$, then it is not necessarily the case that there is a point $\xi \in (0, 1)$ such that $\dot{\gamma}(\xi) = \gamma(1) - \gamma(0)$. Provide a counterexample to substantiate this claim.

However, there is a serviceable substitute for the mean value theorem: if we assume that $\gamma : [a, b] \to \mathbb{R}^N$ is continuously differentiable on an open interval that contains $[a, b]$ and if $M = \max_{t \in [a,b]} |\dot{\gamma}(t)|$ then

$$|\gamma(b) - \gamma(a)| \le M \cdot |b - a|.$$

Prove this statement.

12.3   The Inverse and Implicit Function Theorems

It is easy to tell whether a continuous function of one real variable is invertible. If the function is strictly monotone increasing or strictly monotone decreasing on an interval then the restriction of the function to that interval is invertible. The converse is true as well. It is more difficult to tell whether a function of several variables, when restricted to a neighborhood of a point, is invertible. The reason, of course, is that such a function will in general have different monotonicity behavior in different directions. Also the domain of the function could have a strange shape.

However, if we look at the one-variable situation in a new way, it can be used to give us an idea for analyzing functions of several variables. Suppose
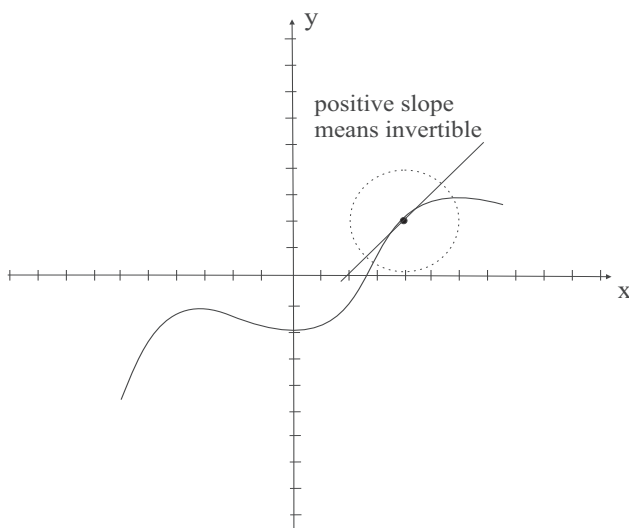
Figure 12.8: An invertible function.

that f is continuously differentiable on an open interval I and that $P \in I$. If $f'(P) > 0$ then the continuity of f' tells us that, for x near P, $f'(x) > 0$. Thus f is strictly increasing on some (possibly smaller) open interval J centered at P. Such a function, when restricted to J, is an invertible function. The same analysis applies when $f'(P) < 0$.

Now the hypothesis that $f'(P) > 0$ or $f'(P) < 0$ has an important geometric interpretation—the positivity of $f'(P)$ means that the tangent line to the graph of f at P has positive slope, hence that the tangent line is the graph of an invertible function (Figure 12.8); likewise the negativity of $f'(P)$ means that the tangent line to the graph of f at P has negative slope, hence that the tangent line is the graph of an invertible function (Figure 12.9). Since the tangent line is a very close approximation at P to the graph of f, our geometric intuition suggests that the local invertibility of f is closely linked to the invertibility of the function describing the tangent line. This guess is in fact borne out in the discussion in the last paragraph.

We would like to carry out an analysis of this kind for a function f from a subset of $\mathbb{R}^k$ into $\mathbb{R}^k$. If P is in the domain of f and if a certain derivative of f at P (to be discussed below) does not vanish, then we would like to conclude that there is a neighborhood U of P such that the restriction of f to U is invertible. That is the content of the Inverse Function Theorem.

Before we formulate and prove this important theorem, we first discuss the kind of derivative of f at P that we shall need to examine.
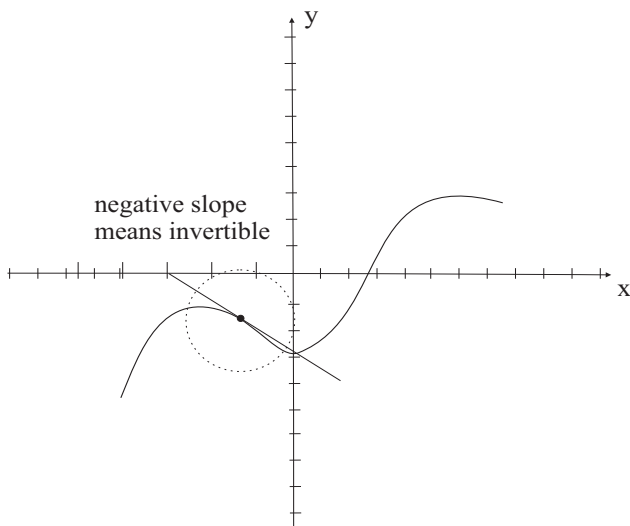
Figure 12.9: Negative slope implies invertible.

Definition 12.21 Let f be a differentiable function from an open subset U of $\mathbb{R}^k$ into $\mathbb{R}^k$. The *Jacobian matrix* of f at a point $\mathbf{P} \in U$ is the matrix

$$Jf(\mathbf{P}) = \begin{pmatrix} \dfrac{\partial f_1}{\partial x_1}(\mathbf{P}) & \dfrac{\partial f_1}{\partial x_2}(\mathbf{P}) & \cdots & \dfrac{\partial f_1}{\partial x_k}(\mathbf{P}) \\[2mm] \dfrac{\partial f_2}{\partial x_1}(\mathbf{P}) & \dfrac{\partial f_2}{\partial x_2}(\mathbf{P}) & \cdots & \dfrac{\partial f_2}{\partial x_k}(\mathbf{P}) \\[2mm] & \cdots & \\[2mm] \dfrac{\partial f_k}{\partial x_1}(\mathbf{P}) & \dfrac{\partial f_k}{\partial x_2}(\mathbf{P}) & \cdots & \dfrac{\partial f_k}{\partial x_k}(\mathbf{P}) \end{pmatrix}.$$

We have seen the matrix in the definition before: It is just the derivative of f at P.

EXAMPLE 12.22 Let $f(x, y) = (x^2 - y, y^2 - x)$. Let $\mathbf{P} = (1, 1)$. Then

$$Jf(\mathbf{P}) = \begin{pmatrix} 2x & -1 \\ -1 & 2y \end{pmatrix}\Bigg|_{\mathbf{P}} = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}.$$

This is the Jacobian matrix for f at $\mathbf{P}$. $\square$

Notice that if we were to expand the function f in a Taylor series about $\mathbf{P}$ (this would be in fact a k-tuple of expansions, since $f = (f_1, f_2, \ldots, f_k)$) then the expansion would be

$$f(\mathbf{P} + \mathbf{h}) = f(\mathbf{P}) + Jf(P)\mathbf{h} + \ldots.$$

Thus the Jacobian matrix is a natural object to study. Moreover we see that the expression $f(\mathbf{P} + \mathbf{h}) - f(\mathbf{P})$ is well approximated by the expression $Jf(\mathbf{P})\mathbf{h}$. Thus, in analogy with one-variable analysis, we might expect that the invertibility of the matrix $Jf(\mathbf{P})$ would imply the existence of a neighborhood of $\mathbf{P}$ on which the function f is invertible. This is indeed the case:

Theorem 12.23 (The Inverse Function Theorem) *Let f be a continuously differentiable function from an open set $U \subseteq \mathbb{R}^k$ into $\mathbb{R}^k$. Suppose that $\mathbf{P} \in U$ and that the matrix $Jf(\mathbf{P})$ is invertible. Then there is a neighborhood V of P such that the restriction of f to V is invertible.*

Proof: The proof of the theorem as stated is rather difficult. Therefore we shall content ourselves with the proof of a special case: we shall make the additional hypothesis that the function f is twice continuously differentiable in a neighborhood of $\mathbf{P}$.

Choose $s > 0$ such that $\overline{B}(\mathbf{P}, s) \subseteq U$ and so that $\det Jf(x) \neq 0$ for all $x \in \overline{B}(P, s)$. Thus the Jacobian matrix $Jf(x)$ is invertible for all $x \in \overline{B}(P, s)$. With the extra hypothesis, Taylor's theorem tells us that there is a constant C such that if $\|\mathbf{h}\| < s/2$ then

$$f(\mathbf{Q} + \mathbf{h}) - f(\mathbf{Q}) = Jf(\mathbf{Q})\mathbf{h} + \mathcal{R}_{1,\mathbf{Q}}(f, \mathbf{h}), \qquad (12.23.1)$$

where

$$|\mathcal{R}_{1,\mathbf{Q}}(\mathbf{h})| \leq C \cdot \frac{\|\mathbf{h}\|^2}{2!},$$

and

$$C = \sup_{\substack{\mathbf{t} \in B(\mathbf{Q}, r) \\ j_1 + j_2 + \cdots + j_k = 2}} \left| \frac{\partial^{j_1 + j_2 + \cdots + j_k} f}{\partial x_1^{j_1} \partial x_2^{j_2} \cdots \partial x_k^{j_k}} \right|.$$

However, all the derivatives in the sum specifying C are, by hypothesis, continuous functions. Since all the balls $B(\mathbf{Q}, s/2)$ are contained in the compact subset $\overline{B}(P, s)$ of U it follows that we may choose C to be a finite number *independent of* $\mathbf{Q}$.

Now the matrix $Jf(\mathbf{Q})^{-1}$ exists by hypothesis. The coefficients of this matrix will be continuous functions of $\mathbf{Q}$ because those of Jf are. Thus these coefficients will be bounded above on $\overline{B}(\mathbf{P}, s)$. We conclude that there is a constant $K > 0$ *independent of* $\mathbf{Q}$ such that for every $\mathbf{k} \in \mathbb{R}^k$ we have

$$\|Jf(\mathbf{Q})^{-1}\mathbf{k}\| \leq K\|\mathbf{k}\|.$$

Taking $\mathbf{k} = Jf(\mathbf{Q})\mathbf{h}$ yields

$$\|\mathbf{h}\| \leq K\|Jf(\mathbf{Q})\mathbf{h}\|. \qquad (12.23.2)$$

Now set

$$r = \min\{s/2, 1/(KC)\}.$$

Line (12.23.1) tells us that, for $\mathbf{Q} \in B(\mathbf{P}, r)$ and $\|\mathbf{h}\| < r$,

$$\|f(\mathbf{Q} + \mathbf{h}) - f(\mathbf{Q})\| \geq \|Jf(Q)\mathbf{h}\| - \|\mathcal{R}_{1,\mathbf{Q}}(\mathbf{h})\|.$$

But estimate (12.23.2), together with our estimate from above on the error term $\mathcal{R}$, yields that the right side of this equation is

$$\geq \frac{\|\mathbf{h}\|}{K} - \frac{C}{2}\|\mathbf{h}\|^2.$$

The choice of r tells us that $\|\mathbf{h}\| \leq 1/(KC)$ hence the last line majorizes $(K/2)\|\mathbf{h}\|$.

But this tells us that, for any $\mathbf{Q} \in B(\mathbf{P}, r)$ and any $\mathbf{h}$ satisfying $\|\mathbf{h}\| < r$, it holds that $f(\mathbf{Q} + \mathbf{h}) \neq f(\mathbf{Q})$. In particular, the function f is one-to-one when restricted to the ball $B(\mathbf{P}, r/2)$. Thus $f|_{B(P,s/2)}$ is invertible. $\square$

In fact the estimate

$$\|f(\mathbf{Q} + \mathbf{h}) - f(\mathbf{Q})\| \geq \frac{K}{2}\|\mathbf{h}\|$$

that we derived easily implies that the image of every $B(\mathbf{Q}, s)$ contains an open ball $B(f(\mathbf{Q}), s')$, some $s' > 0$. This means that f is an *open mapping*. You will be asked in the exercises to provide the details of this assertion.

EXAMPLE 12.24 Let $f(x, y) = (xy - y^3, y + x^2)$. Notice that the Jacobian matrix of this function is

$$Jf = \begin{pmatrix} y & x - 3y^2 \\ 2x & 1 \end{pmatrix}.$$

At the point $(1, 1)$ the Jacobian is

$$Jf(1, 1) = \begin{pmatrix} 1 & -2 \\ 2 & 1 \end{pmatrix}.$$

The determinant of $Jf(1, 1)$ is $5 \neq 0$. Thus the Inverse Function Theorem guarantees that f is invertible in a neighborhood of the point $(1, 1)$. $\square$

EXAMPLE 12.25 Define

$$f(x, y) = (xy + y, y - x).$$

It is easy to calculate that the Jacobian determinant at the point $(1, 1)$ is $3 \neq 0$. So the Inverse Function Theorem applies and we know that f is invertible in a neighborhood.

In this example it is actually possible to calculate f^{-1}, and we ask you to perform this calculation as an exercise. $\square$

With some additional effort it can be shown that f^{-1} is continuously differentiable in a neighborhood of $f(\mathbf{P})$. However, the details of this matter are beyond the scope of this book. We refer the interested reader to [RUD1].

Next we turn to the Implicit Function Theorem. This result addresses the question of when we can solve an equation

$$f(x_1, x_2, \ldots, x_k) = 0$$

for one of the variables in terms of the other $(k-1)$ variables. It is illustrative to first consider a simple example. Look at the equation

$$f(x_1, x_2) = (x_1)^2 + (x_2)^2 = 1.$$

We may restrict attention to $-1 \leq x_1 \leq 1, -1 \leq x_2 \leq 1$. As a glance at the graph shows, we can solve this equation for x_2, uniquely in terms of x_1, in a neighborhood of any point *except* for the points $(\pm 1, 0)$. At these two exceptional points it is impossible to avoid the ambiguity in the square root process, even by restricting to a very small neighborhood. At other points, we may write

$$t_2 = +\sqrt{1 - (t_1)^2}$$

for points (t_1, t_2) near (x_1, x_2) when $x_2 > 0$ and

$$t_2 = -\sqrt{1 - (t_1)^2}$$

for points (t_1, t_2) near (x_1, x_2) when $x_2 < 0$.

What distinguishes the two exceptional points from the others is that the tangent line to the locus (a circle) is vertical at each of these points. Another way of saying this is that

$$\frac{\partial f}{x_2} = 0$$

at these points (Figure 12.10). These preliminary considerations motivate the following theorem.

Theorem 12.26 (The Implicit Function Theorem) *Let f be a function of k real variables, taking scalar values, whose domain contains a neighborhood of a point $\mathbf{P}$. Assume that f is continuously differentiable and that $f(\mathbf{P}) = 0$. If $(\partial f / \partial x_k)(\mathbf{P}) \neq 0$ then there are numbers $\delta > 0, \eta > 0$ such that if $|x_1 - P_1| < \delta$, $|x_2 - P_2| < \delta, \ldots, |x_{k-1} - P_{k-1}| < \delta$, then there is a unique x_k with $|x_k - P_k| < \eta$ and*

$$f(x_1, x_2, \ldots, x_k) = 0. \tag{12.26.1}$$

In other words, in a neighborhood of $\mathbf{P}$, the equation (12.26.1) uniquely determines x_k in terms of $x_1, x_2, \ldots, x_{k-1}$.

Proof: We consider the function

$$T : (x_1, x_2, \ldots, x_k) \longmapsto (x_1, x_2, \ldots, x_{k-1}, f(x_1, x_2, \ldots, x_k)).$$
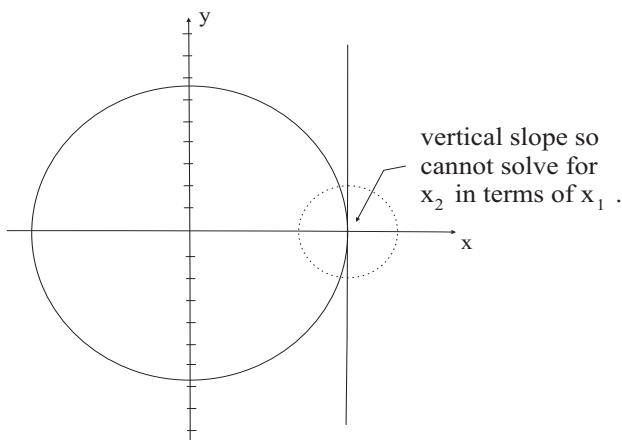
Figure 12.10: Vertical tangents.

The Jacobian matrix of T at $\mathbf{P}$ is

$$
\begin{pmatrix}
1 & 0 & \cdots & 0 \\
0 & 1 & \cdots & 0 \\
& \cdots & & \\
0 & \cdots & 1 & 0 \\
\frac{\partial f}{\partial x_1}(\mathbf{P}) & \frac{\partial f}{\partial x_2}(\mathbf{P}) & \cdots & \frac{\partial f}{\partial x_k}(\mathbf{P})
\end{pmatrix}.
$$

Of course the determinant of this matrix is $\partial f/\partial x_k(\mathbf{P})$, which we hypothesized to be nonzero. Thus the Inverse Function Theorem applies to T. We conclude that T is invertible in a neighborhood of $\mathbf{P}$. That is, there is a number $\eta > 0$ and a neighborhood W of the point $(P_1, P_2, \ldots, P_{k-1}, 0)$ such that

$$ T : B(\mathbf{P}, \eta) \longmapsto W $$

is a one-to-one, onto, continuously differentiable function which is invertible. Select $\delta > 0$ such that if $|x_1 - P_1| < \delta$, $|x_2 - P_2| < \delta$, $\ldots$, $|x_{k-1} - P_{k-1}| < \delta$ then the point $(x_1, x_2, \ldots, x_{k-1}, 0) \in W$. Such a point $(x_1, x_2, \ldots, x_{k-1}, 0)$ then has a unique inverse image under T that lies in $B(\mathbf{P}, \eta)$. But this just says that there is a unique x_k such that $f(x_1, x_2, \ldots, x_k) = 0$. We have established the existence of δ and η as required, hence the proof is complete. $\qquad \square$

EXAMPLE 12.27 Let $f(x, y) = yx^2 - x + y$. Observe that $f(0, 0) = 0$. Also note that

$$ \frac{\partial f}{\partial y}(0, 0) = 1 \neq 0\,. $$

Thus the implicit function theorem guarantees that we can solve for y in terms of x in a neighborhood of the origin. And in fact, in this simple instance, we can solve explicitly:

$$y = \frac{x}{x^2 + 1} \, .$$

□

Exercises

1. Prove that a function satisfying the hypotheses of the Inverse Function Theorem is an open mapping in a neighborhood of the point **P**.

2. Prove that the Implicit Function Theorem is still true if the equation $f(x_1, x_2, \ldots, x_k) = 0$ is replaced by $f(x_1, x_2, \ldots, x_k) = c$. (**Hint:** Do *not* repeat the proof of the Implicit Function Theorem.)

3. Let $y = \varphi(x)$ be a twice continuously differentiable function on $[0, 1]$ with nonvanishing first derivative. Let $\mathcal{U}$ be the graph of φ. Show that there is an open neighborhood W of $\mathcal{U}$ so that, if $P \in W$, then there is a unique point $X \in \mathcal{U}$ which is nearest to P.

4. Give an example of a curve which is not twice continuously differentiable for which the result of Exercise **3** fails.

5. Use the Implicit Function Theorem to show that the natural logarithm function can have only one zero.

6. Use the Implicit Function Theorem to show that the exponential function can have no zeros.

7. It is not true that if a function f from $\mathbb{R}^k$ to $\mathbb{R}^k$ is invertible in a neighborhood of a point P in the domain then the Jacobian determinant at P is nonzero. Provide an example to illustrate this point.

8. It is not true that if the equation $f(x, y) = 0$ can be solved for y in terms of x near the point P then $\partial f / \partial y(P) \neq 0$. Provide an example to illustrate this point.

* 9. Use the Implicit Function Theorem to give a proof of the Fundamental Theorem of Algebra.

Appendix I: Elementary Number Systems

Section A1.1. The Natural Numbers

Mathematics deals with a variety of number systems. The simplest number system is $\mathbb{N}$, the *natural numbers*. As we have already noted, this is just the set of positive integers $\{1, 2, 3, \dots\}$. In a rigorous course of logic, the set $\mathbb{N}$ is constructed from the axioms of set theory. However, in this book we shall assume that you are familiar with the positive integers and their elementary properties.

The principal properties of $\mathbb{N}$ are as follows:

1. 1 is a natural number.

2. If n is a natural number then there is another natural number $\widehat{n}$ which is called the *successor* of n. [We think of the successor of n as the number that comes after n.]

3. $1 \neq \widehat{n}$ for every natural number n.

4. If $\widehat{m} = \widehat{n}$ then $m = n$.

5. *(Principle of Induction)* If $Q(n)$ is a property of the natural number n and if

   (a) The property $Q(1)$ holds;
   (b) Whenever $Q(j)$ holds then it follows that $Q(\widehat{j})$ holds;

   then all natural numbers have the property Q.

These rules, or *axioms*, are known as the *Peano Axioms* for the natural numbers (named after Giuseppe Peano (1858–1932) who developed them). We take it for granted that the usual set of natural numbers satisfies these rules. We see that 1 is in that set. Each positive integer has a "successor"—after 1 comes 2 and after 2 comes 3 and so forth. The number 1 is not the successor of any other positive integer. Two positive integers with the same successor

must be the same. The last axiom is more subtle but makes good sense: if some property $Q(n)$ holds for $n = 1$ and if whenever it holds for n then it also holds for $n + 1$, then we may conclude that Q holds for all positive integers.

We will spend the remainder of this section exploring Axiom **(5)**, the Principle of Induction.

Example A1.1

Let us prove that, for each positive integer n, it holds that

$$1 + 2 + \cdots + n = \frac{n \cdot (n + 1)}{2}.$$

We denote this equation by $Q(n)$, and follow the scheme of the Principle of Induction.

First, $Q(1)$ is true since then both the left and the right side of the equation equal 1. Now assume that $Q(n)$ is true for some natural number n. Our job is to show that it follows that $Q(n + 1)$ is true.

Since $Q(n)$ is true, we know that

$$1 + 2 + \cdots + n = \frac{n \cdot (n + 1)}{2}.$$

Let us add the quantity $n + 1$ to both sides. Thus

$$1 + 2 + \cdots + n + (n + 1) = \frac{n \cdot (n + 1)}{2} + (n + 1).$$

The right side of this new equality simplifies and we obtain

$$1 + 2 + \cdots + (n + 1) = \frac{(n + 1) \cdot ((n + 1) + 1)}{2}.$$

But this is just $Q(n+1)$ or $Q(\widehat{n})$! *We have assumed $Q(n)$ and have proved $Q(\widehat{n})$,* just as the Principle of Induction requires.

Thus we may conclude that property Q holds for all positive integers, as desired.   □

The formula that we derived in Example A1.1 was probably known to the ancient Greeks. However, a celebrated anecdote credits Carl Friedrich Gauss (1777–1855) with discovering the formula when he was nine years old. Gauss went on to become (along with Isaac Newton and Archimedes) one of the three greatest mathematicians of all time.

The formula from Example A1.1 gives a neat way to add up the integers from 1 to n, for any n, without doing any work. Any time that we discover a new mathematical fact, there are generally several others hidden within it. The next example illustrates this point.

Example A1.2

The sum of the first m positive even integers is $m \cdot (m+1)$. To see this note that the sum in question is

$$2 + 4 + 6 + \cdots + 2m = 2(1 + 2 + 3 + \cdots + m).$$

But, by the first example, the sum in parentheses on the right is equal to $m \cdot (m+1)/2$. It follows that

$$2 + 4 + 6 + \cdots + 2m = 2 \cdot \frac{m \cdot (m+1)}{2} = m \cdot (m+1).$$

$\square$

The second example could also be performed by induction (without using the result of the first example).

Example A1.3

Now we will use induction incorrectly to prove a statement that is completely preposterous:

All horses are the same color.

There are finitely many horses in existence, so it is convenient for us to prove the slightly more technical statement

Any collection of k horses consists of horses which are all the same color.

Our statement $Q(k)$ is this last displayed statement.

Now $Q(1)$ is true: *one horse is the same color.* (Note: this is not a joke, and the error has not occurred yet.)

Suppose next that $Q(k)$ is true: we assume that any collection of k horses has the same color. Now consider a collection of $\widehat{k} = k + 1$ horses. Remove one horse from that collection. By our hypothesis, the remaining k horses have the same color.

Now replace the horse that we removed and remove a different horse. Again, the remaining k horses have the same color.

We keep repeating this process: remove each of the $k + 1$ horses one by one and conclude that the remaining k horses have the same color. Therefore every horse in the collection is the same color as every other. So all $k + 1$ horses have the same color. The statement $Q(k + 1)$ is thus proved (assuming the truth of $Q(k)$) and the induction is complete.

Where is our error? It is nothing deep—just an oversight. The argument we have given is wrong when $\widehat{k} = k + 1 = 2$. For remove one horse from a set of two and the remaining (*one*) horse is the same color. Now replace the removed

horse and remove the other horse. The remaining (*one*) horse is the same color. *So what?* We cannot conclude that the two horses are colored the same. Thus the induction breaks down at the outset; the reasoning is incorrect.   □

Proposition A1.4

Let a and b be real numbers and n a natural number. Then

$$
\begin{aligned}
(a+b)^n \;=\; & a^n + \frac{n}{1}a^{n-1}b + \frac{n(n-1)}{2\cdot 1}a^{n-2}b^2 \\
& + \frac{(n(n-1)(n-2)}{3\cdot 2\cdot 1}a^{n-3}b^3 \\
& + \cdots + \frac{n(n-1)\cdots 2}{(n-1)(n-2)\cdots 2\cdot 1}ab^{n-1} + b^n.
\end{aligned}
$$

Proof: The case $n = 1$ being obvious, proceed by induction.   □

Example A1.5

The expression

$$
\frac{n(n-1)\cdots(n-k+1)}{k(k-1)\cdots 1}
$$

is often called the *kth binomial coefficient* and is denoted by the symbol

$$
\left(\begin{array}{c} n \\ k \end{array} \right).
$$

Using the notation $m! = m \cdot (m-1) \cdot (m-2) \cdots 2 \cdot 1$, for m a natural number, we may write the kth binomial coefficient as

$$
\left(\begin{array}{c} n \\ k \end{array} \right) = \frac{n!}{(n-k)! \cdot k!}.
$$ □

Section A1.2. The Integers

Now we will apply the notion of an equivalence class (see Appendix II) to *construct* the integers (both positive and negative). There is an important point of knowledge to be noted here. For the sake of having a reasonable place to begin our work, we took the natural numbers $\mathbb{N} = \{1, 2, 3, \dots\}$ as given. Since the natural numbers have been used for thousands of years to keep track of objects for barter, this is a plausible thing to do. Even people who know no mathematics accept the positive integers. However, the number zero and the negative numbers are a different matter. It was not until the fifteenth century

that the concepts of zero and negative numbers started to take hold—for they do not correspond to explicit collections of objects (five fingers or ten shoes) but rather to *concepts* (zero books is the lack of books; minus 4 pens means that we owe someone four pens). After some practice we get used to negative numbers, but explaining in words what they mean is always a bit clumsy.

It is much more satisfying, from the point of view of logic, to *construct* the integers (including the negative whole numbers and zero) from what we already have, that is, from the natural numbers. We proceed as follows. Let $A = \mathbb{N} \times \mathbb{N}$, the set of ordered pairs of natural numbers. We define a relation (see Appendix II, Section 6) $\mathcal{R}$ on A and A as follows:

$$(a, b) \text{ is related to } (a', b') \text{ if } a + b' = a' + b$$

See also Appendix II, Section 6 for the concept of equivalence relation.

Theorem A1.6

The relation $\mathcal{R}$ is an equivalence relation.

Proof: That (a, b) is related to (a, b) follows from the trivial identity $a + b = a + b$. Hence $\mathcal{R}$ is reflexive. Second, if (a, b) is related to (a', b') then $a + b' = a' + b$ hence $a' + b = a + b'$ (just reverse the equality) hence (a', b') is related to (a, b). So $\mathcal{R}$ is symmetric.

Finally, if (a, b) is related to (a', b') and (a', b') is related to (a'', b'') then we have

$$a + b' = a' + b \qquad \text{and} \qquad a' + b'' = a'' + b'.$$

Adding these equations gives

$$(a + b') + (a' + b'') = (a' + b) + (a'' + b').$$

Cancelling a' and b' from each side finally yields

$$a + b'' = a'' + b.$$

Thus (a, b) is related to (a'', b''). Therefore $\mathcal{R}$ is transitive. We conclude that $\mathcal{R}$ is an equivalence relation. $\square$

Now our job is to understand the equivalence classes which are induced by $\mathcal{R}$. [We will ultimately call this number system the integers $\mathbb{Z}$.] Let $(a, b) \in A$ and let $[(a, b)]$ be the corresponding equivalence class. If $b > a$ then we will denote this equivalence class by the integer $b - a$. For instance, the equivalence class $[(2, 7)]$ will be denoted by 5. Notice that if $(a', b') \in [(a, b)]$ then $a + b' = a' + b$ hence $b' - a' = b - a$. Therefore the integer symbol that we choose to represent our equivalence class is *independent of which element of the equivalence class is used to compute it.*

If $(a, b) \in A$ and $b = a$ then we let the symbol 0 denote the equivalence class $[(a, b)]$. Notice that if (a', b') is any other element of $[(a, b)]$ then it must be that $a + b' = a' + b$ hence $b' = a'$; therefore this definition is unambiguous.

If $(a, b) \in A$ and $a > b$ then we will denote the equivalence class $[(a, b)]$ by the symbol $-(a - b)$. For instance, we will denote the equivalence class $[(7, 5)]$ by the symbol -2. Once again, if (a', b') is related to (a, b) then the equation $a + b' = a' + b$ guarantees that our choice of symbol to represent $[(a, b)]$ is unambiguous.

Thus we have given our equivalence classes names, and these names *look just like* the names that we usually give to integers: there are positive integers, and negative ones, and zero. But we want to see that these objects *behave* like integers. (As you read on, use the intuitive, non-rigorous mnemonic that the equivalence class $[(a, b)]$ stands for the integer $b - a$.)

First, do these new objects that we have constructed *add* correctly? Well, let $X = [(a, b)]$ and $Y = [(c, d)]$ be two equivalence classes. *Define* their sum to be $X + Y = [(a + c, b + d)]$. We must check that this is unambiguous. If $(\widetilde{a}, \widetilde{b})$ is related to (a, b) and $(\widetilde{c}, \widetilde{d})$ is related to (c, d) then of course we know that

$$a + \widetilde{b} = \widetilde{a} + b$$

and

$$c + \widetilde{d} = \widetilde{c} + d\,.$$

Adding these two equations gives

$$(a + c) + (\widetilde{b} + \widetilde{d}) = (\widetilde{a} + \widetilde{c}) + (b + d)$$

hence $(a+c, b+d)$ is related to $(\widetilde{a}+\widetilde{c}, \widetilde{b}+\widetilde{d})$. Thus, adding two of our equivalence classes gives another equivalence class, as it should.

Example A1.7

To add 5 and 3 we first note that 5 is the equivalence class $[(2, 7)]$ and 3 is the equivalence class $[(2, 5)]$. We add them componentwise and find that the sum is $[(2 + 2, 7 + 5)] = [(4, 12)]$. Which equivalence class is this answer? Looking back at our prescription for giving names to the equivalence classes, we see that this is the equivalence class that we called $12 - 4$ or 8. So we have rediscovered the fact that $5 + 3 = 8$. Check for yourself that if we were to choose a different representative for 5—say $(6, 11)$—and a different representative for 3—say $(24, 27)$—then the same answer would result.

Now let us add 4 and -9. The first of these is the equivalence class $[(3, 7)]$ and the second is the equivalence class $[(13, 4)]$. The sum is therefore $[(16, 11)]$, and this is the equivalence class that we call $-(16 - 11)$ or -5. That is the answer that we would expect when we add 4 to -9.

Next, we add -12 and -5. Previous experience causes us to expect the answer to be -17. Now -12 is the equivalence class $[(19, 7)]$ and -5 is the

equivalence class $[(7, 2)]$. The sum is $[(26, 9)]$, which is the equivalence class that we call -17.

Finally, we can see in practice that our method of addition is unambiguous. Let us redo the second example using $[(6, 10)]$ as the equivalence class represented by 4 and $[(15, 6)]$ as the equivalence class represented by -9. Then the sum is $[(21, 16)]$, and this is still the equivalence class -5, as it should be. $\square$

The assertion that the result of calculating a sum—no matter which representatives we choose for the equivalence classes—will give only one answer is called the "fact that addition is *well defined*." In order for our definitions to make sense, it is essential that we check this property of well-definedness.

Remark A1.8

What is the point of this section? Everyone knows about negative numbers, so why go through this abstract construction? The reason is that, until one sees this construction, negative numbers are just imaginary objects—placeholders if you will—which are a useful notation but which do not exist. Now they *do* exist. They are a collection of equivalence classes of pairs of natural numbers. This collection is equipped with certain arithmetic operations, such as addition, subtraction, and multiplication. We now discuss these last two.

If $x = [(a, b)]$ and $y = [(c, d)]$ are integers, we define their *difference* to be the equivalence class $[(a + d, b + c)]$; we denote this difference by $x - y$.

Example A1.9

We calculate $8 - 14$. Now $8 = [(1, 9)]$ and $14 = [(3, 17)]$. Therefore

$$8 - 14 = [(1 + 17, 9 + 3)] = [(18, 12)] = -6\,,$$

as expected.

As a second example, we compute $(-4) - (-8)$. Now

$$-4 - (-8) = [(6, 2)] - [(13, 5)] = [(6 + 5, 2 + 13)] = [(11, 15)] = 4\,.$$

Remark A1.10

When we first learn that $(-4) - (-8) = (-4) + 8 = 4$, the explanation is a bit mysterious: why is "minus a minus equal to a plus"? Now there is no longer any mystery: this property follows *from our construction* of the number system $\mathbb{Z}$. $\square$

Finally, we turn to multiplication. If $x = [(a, b)]$ and $y = [(c, d)]$ are integers then we define their product by the formula

$$x \cdot y = [(a \cdot d + b \cdot c, a \cdot c + b \cdot d)].$$

This definition may be a surprise. Why did we not define $x \cdot y$ to be $[(a \cdot c, b \cdot d)]$? There are several reasons: first of all, the latter definition would give the wrong answer; moreover, it is not unambiguous (different representatives of x and y would give a different answer). If you recall that we think of $[(a, b)]$ as representing $b - a$ and $[(c, d)]$ as representing $d - c$ then the product should be the equivalence class that represents $(b - a) \cdot (d - c)$. That is the motivation behind our definition.

We proceed now to an example.

Example A1.11

We compute the product of -3 and -6. Now

$$(-3) \cdot (-6) = [(5, 2)] \cdot [(9, 3)] = [(5 \cdot 3 + 2 \cdot 9, 5 \cdot 9 + 2 \cdot 3)] = [(33, 51)] = 18 \,,$$

which is the expected answer.

As a second example, we multiply -5 and 12. We have

$$-5 \cdot 12 = [(7, 2)] \cdot [(1, 13)] = [(7 \cdot 13 + 2 \cdot 1, 7 \cdot 1 + 2 \cdot 13)] = [(93, 33)] = -60 \,.$$

Finally, we show that 0 times any integer A equals zero. Let $A = [(a, b)]$. Then

$$\begin{aligned} 0 \cdot A &= [(1, 1)] \cdot [(a, b)] = [(1 \cdot b + 1 \cdot a, 1 \cdot a + 1 \cdot b)] \\ &= [(a + b, a + b)] \\ &= 0 \,. \end{aligned} \qquad \square$$

Remark A1.12

Notice that one of the pleasant by-products of our construction of the integers is that we no longer have to give artificial explanations for why the product of two negative numbers is a positive number or why the product of a negative number and a positive number is negative. These properties instead follow automatically from our construction.

Of course we will not discuss division for integers; in general division of one integer by another makes no sense *in the universe of the integers*.

In the rest of this book we will follow the standard mathematical custom of denoting the set of all integers by the symbol $\mathbb{Z}$. We will write the integers

not as equivalence classes, but in the usual way as $\cdots - 3, -2, -1, 0, 1, 2, 3, \ldots$. The equivalence classes are a device that we used to *construct* the integers. Now that we have the integers in hand, we may as well write them in the simple, familiar fashion.

In an exhaustive treatment of the construction of $\mathbb{Z}$, we would prove that addition and multiplication are commutative and associative, prove the distributive law, and so forth. But the purpose of this section is to demonstrate modes of logical thought rather than to be thorough.

Section A1.3. The Rational Numbers

In this section we use the integers, together with a construction using equivalence classes, to build the rational numbers. Let A be the set $\mathbb{Z} \times (\mathbb{Z} \setminus \{0\})$. Here the symbol $\setminus$ stands for "subtraction of sets": $\mathbb{Z} \setminus \{0\}$ denotes the set of all elements of $\mathbb{Z}$ *except* 0. In other words, A is the set of ordered pairs (a, b) of integers subject to the condition that $b \neq 0$. [*Think, intuitively and non-rigorously, of this ordered pair as "representing" the fraction a/b.*] We definitely want it to be the case that certain ordered pairs represent the same number. For instance,

The number $\frac{1}{2}$ should be the same number as $\frac{3}{6}$.

This example motivates our equivalence relation. Declare (a, b) to be related to (a', b') if $a \cdot b' = a' \cdot b$. [*Here we are thinking, intuitively and non-rigorously, that the fraction a/b should equal the fraction a'/b' precisely when $a \cdot b' = a' \cdot b$.*]

Is this an equivalence relation? Obviously the pair (a, b) is related to itself, since $a \cdot b = a \cdot b$. Also the relation is symmetric: if (a, b) and (a', b') are pairs and $a \cdot b' = a' \cdot b$ then $a' \cdot b = a \cdot b'$. Finally, if (a, b) is related to (a', b') and (a', b') is related to (a'', b'') then we have both

$$a \cdot b' = a' \cdot b \quad \text{and} \quad a' \cdot b'' = a'' \cdot b'.$$

Multiplying the left sides of these two equations together and the right sides together gives

$$(a \cdot b') \cdot (a' \cdot b'') = (a' \cdot b) \cdot (a'' \cdot b').$$

If $a' = 0$ then it follows immediately that both a and a'' must be zero. So the three pairs $(a, b), (a', b')$, and (a'', b'') are equivalent and there is nothing to prove. So we may assume that $a' \neq 0$. We know *a priori* that $b' \neq 0$; therefore we may cancel common terms in the last equation to obtain

$$a \cdot b'' = b \cdot a''.$$

Thus (a, b) is related to (a'', b''), and our relation is transitive.

The resulting collection of equivalence classes will be called the set of *rational numbers*, and we shall denote this set with the symbol $\mathbb{Q}$.

Example A1.13

The equivalence class $[(4, 12)]$ in the rational numbers contains all of the pairs $(4, 12), (1, 3), (-2, -6)$. (Of course it contains infinitely many other pairs as well.) This equivalence class represents the fraction $4/12$, which we sometimes also write as $1/3$ or $-2/(-6)$.   □

If $[(a, b)]$ and $[(c, d)]$ are rational numbers then we define their *product* to be the rational number

$$[(a \cdot c, b \cdot d)] \, .$$

This is well defined, for if (a, b) is related to $(\widetilde{a}, \widetilde{b})$ and (c, d) is related to $(\widetilde{c}, \widetilde{d})$ then we have the equations

$$a \cdot \widetilde{b} = \widetilde{a} \cdot b \ \text{ and } \ c \cdot \widetilde{d} = \widetilde{c} \cdot d.$$

Multiplying together the left sides and the right sides we obtain

$$(a \cdot \widetilde{b}) \cdot (c \cdot \widetilde{d}) = (\widetilde{a} \cdot b) \cdot (\widetilde{c} \cdot d) \, .$$

Rearranging, we have

$$(a \cdot c) \cdot (\widetilde{b} \cdot \widetilde{d}) = (\widetilde{a} \cdot \widetilde{c}) \cdot (b \cdot d) \, .$$

But this says that the product of $[(a, b)]$ and $[(c, d)]$ is related to the product of $[(\widetilde{a}, \widetilde{b})]$ and $[(\widetilde{c}, \widetilde{d})]$. So multiplication is unambiguous (i.e., well defined).

Example A1.14

The product of the two rational numbers $[(3, 8)]$ and $[(-2, 5)]$ is

$$[(3 \cdot (-2), 8 \cdot 5)] = [(-6, 40)] = [(-3, 20)] \, .$$

This is what we expect: the product of $3/8$ and $-2/5$ is $-3/20$.   □

If $q = [(a, b)]$ and $r = [(c, d)]$ are rational numbers and if r is not zero (that is, $[(c, d)]$ is not the equivalence class zero—in other words, $c \neq 0$) then we define the quotient q/r to be the equivalence class

$$[(ad, bc)] \, .$$

We leave it to you to check that this operation is well defined.

Example A1.15

The quotient of the rational number $[(4, 7)]$ by the rational number $[(3, -2)]$ is, by definition, the rational number

$$[(4 \cdot (-2), 7 \cdot 3)] = [(-8, 21)] \, .$$

This is what we expect: the quotient of $4/7$ by $-3/2$ is $-8/(21)$. □

How should we add two rational numbers? We could try declaring $[(a, b)] + [(c, d)]$ to be $[(a + c, b + d)]$, but this will not work (think about the way that we usually add fractions). Instead we define

$$[(a, b)] + [(c, d)] = [(a \cdot d + c \cdot b, b \cdot d)].$$

We turn now to an example.

Example A1.16

The sum of the rational numbers $[(3, -14)]$ and $[(9, 4)]$ is given by

$$[(3 \cdot 4 + 9 \cdot (-14), (-14) \cdot 4)] = [(-114, -56)] = [(57, 28)].$$

This is consistent with the usual way that we add fractions:

$$-\frac{3}{14} + \frac{9}{4} = \frac{57}{28}.$$ □

Notice that the equivalence class $[(0, 1)]$ is the rational number that we usually denote by 0. It is the additive identity, for if $[(a, b)]$ is another rational number then

$$[(0, 1)] + [(a, b)] = [(0 \cdot b + a \cdot 1, 1 \cdot b)] = [(a, b)].$$

A similar argument shows that $[(0, 1)]$ times any rational number gives $[(0, 1)]$ or 0.

Of course the concept of subtraction is really just a special case of addition (that is $x - y$ is the same thing as $x + (-y)$). So we shall say nothing further about subtraction.

In practice we will write rational numbers in the traditional fashion:

$$\frac{2}{5}, \frac{-19}{3}, \frac{22}{2}, \frac{24}{4}, \ldots.$$

In mathematics it is generally not wise to write rational numbers in mixed form, such as $2\frac{3}{5}$, because the juxtaposition of two numbers could easily be mistaken for multiplication. Instead we would write this quantity as the improper fraction $13/5$.

Definition A1.17

A set S is called a *field* if it is equipped with a binary operation (usually called addition and denoted "+") and a second binary operation (called multiplication and denoted "·") such that the following axioms are satisfied:

A1. S is closed under addition: if $x, y \in S$ then $x + y \in S$.

A2. Addition is commutative: if $x, y \in S$ then $x + y = y + x$.

A3. Addition is associative: if $x, y, z \in S$ then $x + (y + z) = (x + y) + z$.

A4. There exists an element, called 0, in S which is an additive identity: if $x \in S$ then $0 + x = x$.

A5. Each element of S has an additive inverse: if $x \in S$ then there is an element $-x \in S$ such that $x + (-x) = 0$.

M1. S is closed under multiplication: if $x, y \in S$ then $x \cdot y \in S$.

M2. Multiplication is commutative: if $x, y \in S$ then $x \cdot y = y \cdot x$.

M3. Multiplication is associative: if $x, y, z \in S$ then $x \cdot (y \cdot z) = (x \cdot y) \cdot z$.

M4. There exists an element, called 1, which is a multiplicative identity: if $x \in S$ then $1 \cdot x = x$.

M5. Each nonzero element of S has a multiplicative inverse: if $0 \neq x \in S$ then there is an element $x^{-1} \in S$ such that $(x^{-1}) \cdot x = 1$. The element x^{-1} is sometimes denoted $1/x$.

D1. Multiplication distributes over addition: if $x, y, z \in S$ then

$$x \cdot (y + z) = x \cdot y + x \cdot z \,.$$

Eleven axioms is a lot to digest all at once, but in fact these are all familiar properties of addition and multiplication of rational numbers that we use every day: the set $\mathbb{Q}$, with the usual notions of addition and multiplication, forms a field. The integers, by contrast, do not: nonzero elements of $\mathbb{Z}$ (except 1 and -1) do not have multiplicative inverses *in the integers*.

Let us now consider some consequence of the field axioms.

Theorem A1.18

Any field has the following properties:

(1) *If $z + x = z + y$ then $x = y$.*

(2) *If $x + z = 0$ then $z = -x$ (the additive inverse is unique).*

(3) *$-(-y) = y$.*

(4) *If $y \neq 0$ and $y \cdot x = y \cdot z$ then $x = z$.*

(5) *If $y \neq 0$ and $y \cdot z = 1$ then $z = y^{-1}$ (the multiplicative inverse is unique).*

(6) $\left(x^{-1}\right)^{-1} = x.$

(7) $0 \cdot x = 0.$

(8) *If $x \cdot y = 0$ then either $x = 0$ or $y = 0$.*

(9) $(-x) \cdot y = -(x \cdot y) = x \cdot (-y).$

(10) $(-x) \cdot (-y) = x \cdot y.$

Proof: These are all familiar properties of the rationals, but now we are considering them for an arbitrary field. We prove just a few to illustrate the logic.

To prove **(1)** we write

$$z + x = z + y \Rightarrow (-z) + (z + x) = (-z) + (z + y)$$

and now Axiom **A3** yields that this implies

$$((-z) + z) + x = ((-z) + z) + y.$$

Next, Axiom **A5** yields that

$$0 + x = 0 + y$$

and hence, by Axiom **A4**,

$$x = y.$$

To prove **(7)**, we observe that

$$0 \cdot x = (0 + 0) \cdot x,$$

which by Axiom **M2** equals

$$x \cdot (0 + 0).$$

By Axiom **D1** the last expression equals

$$x \cdot 0 + x \cdot 0,$$

which by Axiom **M2** equals $0 \cdot x + 0 \cdot x$. Thus we have derived the equation

$$0 \cdot x = 0 \cdot x + 0 \cdot x.$$

Axioms **A4** and **A2** let us rewrite the left side as

$$0 \cdot x + 0 = 0 \cdot x + 0 \cdot x.$$

Finally, part **(1)** of the present theorem (which we have already proved) yields that

$$0 = 0 \cdot x,$$

which is the desired result.

To prove **(8)**, we suppose that $x \neq 0$. In this case x has a multiplicative inverse x^{-1} and we multiply both sides of our equation by this element:

$$x^{-1} \cdot (x \cdot y) = x^{-1} \cdot 0 \,.$$

By Axiom **M3**, the left side can be rewritten and we have

$$(x \cdot x^{-1}) \cdot y = x^{-1} \cdot 0 \,.$$

Next, we rewrite the right side using Axiom **M2**:

$$(x \cdot x^{-1}) \cdot y = 0 \cdot x^{-1} \,.$$

Now Axiom **M5** allows us to simplify the left side:

$$1 \cdot y = 0 \cdot x^{-1} \,.$$

We further simplify the left side using Axiom **M4** and the right side using Part **(7)** of the present theorem (which we just proved) to obtain:

$$y = 0 \,.$$

Thus we see that if $x \neq 0$ then $y = 0$. But this is logically equivalent with $x = 0$ or $y = 0$, as we wished to prove. [If you have forgotten why these statements are logically equivalent, write a truth table.]   □

Definition A1.19

Let A be a set. We shall say that A is *ordered* if there is a relation $\mathcal{R}$ on A and A satisfying the following properties:

1. If $a \in A$ and $b \in A$ then one and only one of the following holds: $(a, b) \in \mathcal{R}$ or $(b, a) \in \mathcal{R}$ or $a = b$.

2. If a, b, c are elements of A and $(a, b) \in \mathcal{R}$ and $(b, c) \in \mathcal{R}$ then $(a, c) \in \mathcal{R}$.

We call the relation $\mathcal{R}$ an *order* on A.

Rather than write an ordering relation as $(a, b) \in \mathcal{R}$ it is usually more convenient to write it as $a < b$. The notation $b > a$ means the same thing as $a < b$.

Example A1.20

The integers $\mathbb{Z}$ form an ordered set with the usual ordering $<$. We can make this ordering precise by saying that $x < y$ if $y - x$ is a positive integer. For instance,

$$6 < 8 \quad \text{because} \quad 8 - 6 = 2 > 0 \,.$$

Likewise,

$$-5 < -1 \quad \text{because} \quad -1 - (-5) = 4 > 0 \,.$$

Observe that the same ordering works on the rational numbers. □

If A is an ordered set and a, b are elements then we often write $a \leq b$ to mean that *either $a = b$ or $a < b$.*

When a field has an ordering which is compatible with the field operations then a richer structure results:

Definition A1.21

A field F is called an *ordered field* if F has an ordering $<$ that satisfies the following addition properties:

(1) If $x, y, z \in F$ and $y < z$ then $x + y < x + z$.

(2) If $x, y \in F, x > 0$, and $y > 0$ then $x \cdot y > 0$.

Again, these are familiar properties of the rational numbers: $\mathbb{Q}$ forms an ordered field. But there are many other ordered fields as well (for instance, the real numbers $\mathbb{R}$ form an ordered field).

Theorem A1.22

Any ordered field has the following properties:

(1) *If $x > 0$ and $z < y$ then $x \cdot z < x \cdot y$.*

(2) *If $x < 0$ and $z < y$ then $x \cdot z > x \cdot y$.*

(3) *If $x > 0$ then $-x < 0$. If $x < 0$ then $-x > 0$.*

(4) *If $0 < y < x$ then $0 < 1/x < 1/y$.*

(5) *If $x \neq 0$ then $x^2 > 0$.*

(6) *If $0 < x < y$ then $x^2 < y^2$.*

Proof: Again we prove just a few of these statements.

To prove **(1)**, observe that the property **(1)** of ordered fields together with our hypothesis implies that

$$(-z) + z < (-z) + y \,.$$

Thus, using **(A2)**, we see that $y - z > 0$. Since $x > 0$, property **(2)** of ordered fields gives

$$x \cdot (y - z) > 0 \,.$$

Finally,
$$x \cdot y = x \cdot [(y - z) + z] = x \cdot (y - z) + x \cdot z > 0 + x \cdot z$$

(by property **(1)** again). In conclusion,

$$x \cdot y > x \cdot z \,.$$

To prove **(3)**, begin with the equation

$$0 = -x + x \,.$$

Since $x > 0$, the right side is greater than $-x$. Thus $0 > -x$ as claimed. The proof of the other statement of **(3)** is similar.

To prove **(5)**, we consider two cases. If $x > 0$ then $x^2 \equiv x \cdot x$ is positive by property **(2)** of ordered fields. If $x < 0$ then $-x > 0$ (by part **(3)** of the present theorem, which we just proved) hence $(-x) \cdot (-x) > 0$. But part **(10)** of the last theorem guarantees that $(-x) \cdot (-x) = x \cdot x$ hence we see that $x \cdot x > 0$. $\square$

We conclude this Appendix by recording an inadequacy of the field of rational numbers; this will serve in part as motivation for learning about the real numbers in Chapter 1.

Theorem A1.23
   There is no positive rational number q such that $q^2 = q \cdot q = 2$.

Proof: Seeking a contradiction, suppose that there is such a q. Write q in lowest terms as
$$q = \frac{a}{b},$$
with a and b greater than zero. This means that the numbers a and b have no common divisors except 1. The equation $q^2 = 2$ can then be written as

$$a^2 = 2 \cdot b^2 \,.$$

Since 2 divides the right side of this last equation, it follows that 2 divides the left side. But 2 can divide a^2 only if 2 divides a (because 2 is prime). We write $a = 2 \cdot \alpha$ for some positive integer α. But then the last equation becomes

$$4 \cdot \alpha^2 = 2 \cdot b^2 \,.$$

Simplifying yields that
$$2 \cdot \alpha^2 = b^2 \,.$$

Since 2 divides the left side, we conclude that 2 must divide the right side. But 2 can divide b^2 only if 2 divides b.

This is our contradiction: we have argued that 2 divides a *and* that 2 divides b. But a and b were assumed to *have no common divisors*. We conclude that the rational number q cannot exist. $\square$

In fact it turns out that a positive integer can be the square of a rational number if and only if it is the square of a positive integer. This assertion is a special case of a more general phenomenon in number theory known as Gauss's lemma.

Appendix II: Logic and Set Theory

Everyday language is imprecise. Because we are imprecise by *convention*, we can make statements like

<div align="center">

All automobiles are not alike.

</div>

and feel confident that the listener knows that we actually *mean*

<div align="center">

Not all automobiles are alike.

</div>

We can also use spurious reasoning like

<div align="center">

If it's raining then it's cloudy.
It is not raining.
Therefore there are no clouds.

</div>

and not expect to be challenged, because virtually everyone is careless when communicating informally. (Examples of this type will be considered in more detail later.)

Mathematics cannot tolerate this lack of rigor and precision. In order to achieve any depth beyond the most elementary level, we must adhere to strict rules of logic. The purpose of the present Appendix is to discuss the foundations of formal reasoning.

In this chapter we will often use numbers to illustrate logical concepts. The number systems we will encounter are

- The natural numbers $\mathbb{N} = \{1, 2, 3, \dots\}$

- The integers $\mathbb{Z} = \{\dots, -3, -2, -1, 0, 1, 2, 3, \dots\}$

- The rational numbers $\mathbb{Q} = \{p/q : p \text{ is an integer}, q \text{ is an integer}, q \neq 0\}$

- The real numbers $\mathbb{R}$, consisting of all terminating and non-terminating decimal expansions.

[Chapter 1](#) reviewed the real and complex numbers. If you need to review the other number systems, then refer to [Appendix I](#) or look at [KRA1]. For now we assume that you have seen these number systems before. They are convenient for illustrating the logical principles we are discussing.

Section A2.1. "And" and "Or"

The statement

$$\text{"}\mathbf{A} \ \ \text{and} \ \ \mathbf{B}\text{"}$$

means that both **A** is true *and* **B** is true. For instance,

George is tall and George is intelligent.

means both that George is tall *and* George is intelligent. If we meet George and he turns out to be short and intelligent, then the statement is false. If he is tall and stupid then the statement is false. Finally, if George is *both* short and stupid then the statement is false. The statement is *true* precisely when both properties—intelligence and tallness—hold. We may summarize these assertions with a *truth table*. We let

A = George is tall.

and

B = George is intelligent.

The expression

$$\mathbf{A} \wedge \mathbf{B}$$

will denote the phrase "**A** and **B**." In particular, the symbol $\wedge$ is used to denote "and." The letters "T" and "F" denote "True" and "False," respectively. Then we have

A	**B**	**A $\wedge$ B**
T	T	T
T	F	F
F	T	F
F	F	F

Notice that we have listed all possible truth values of **A** and **B** and the corresponding values of the *conjunction* $\mathbf{A} \wedge \mathbf{B}$.

It is a good idea to always list the truth values of **A** and **B** in a truth table in the same way. This will facilitate the comparison and contrast of truth tables.

In a restaurant the menu often contains phrases like

soup or salad

This means that we may select soup *or* select salad, but we may not select both. This use of "or" is called the *exclusive* "or"; it is not the meaning of "or" that we use in mathematics and logic. In mathematics we instead say that "**A** **or** **B**" is true provided that **A** is true or **B** is true or *both* are true. This is the *inclusive* meaning of the word "or." If we let **A** ∨ **B** denote "**A or B**" (the symbol ∨ denotes "or") then the truth table is

A	**B**	**A ∨ B**
T	T	T
T	F	T
F	T	T
F	F	F

The only way that "**A** **or** **B**" can be false is if *both* **A** is false and **B** is false. For instance, the statement

Gary is handsome or Gary is rich.

means that Gary is either handsome or rich or both. In particular, he will not be both ugly and poor. Another way of saying this is that if he is poor he will compensate by being handsome; if he is ugly he will compensate by being rich. *But he could be both handsome and rich.*

We use the inclusive meaning of the word "or" because it gives rise to useful logical equivalences. We treat these later.

Example A2.1

The statement

$$x > 5 \quad \text{and} \quad x < 7$$

is true for the number $x = 11/2$ because this value of x is both greater than 5 *and* less than 7. It is false for $x = 8$ because this x is greater than 5 but not less than 7. It is false for $x = 3$ because this x is less than 7 but not greater than 5. □

Example A2.2

The statement

$$x \text{ is even and } x \text{ is a perfect square}$$

is true for $x = 4$ because both assertions hold. It is false for $x = 2$ because this x, while even, is not a square. It is false for $x = 9$ because this x, while a square, is not even. It is false for $x = 5$ because this x is neither a square nor an even number.

□

Example A2.3

The statement

$$x > 5 \quad \text{or} \quad x \leq 2$$

is true for $x = 1$ since this x is ≤ 2 (even though it is not > 5). It holds for $x = 6$ because this x is > 5 (even though it is not ≤ 2). The statement fails for $x = 3$ since this x is neither > 5 nor ≤ 1. There is no x which is both > 5 and ≤ 2.   □

Example A2.4

The statement

$$x > 5 \quad \text{or} \quad x < 7$$

is true for every real x. For $x = 6$, both statements are true. For $x = 2$, just the second statement is true. For $x = 8$, just the first statement is true.   □

Example A2.5

The statement $(\mathbf{A} \vee \mathbf{B}) \wedge \mathbf{B}$ has the following truth table:

A	B	A ∨ B	(A ∨ B) ∧ B
T	T	T	T
T	F	T	F
F	T	T	T
F	F	F	F

□

The words "and" and "or" are called *connectives*: their role in sentential logic is to enable us to build up (or connect together) pairs of statements. In the next section we will become acquainted with the other two basic connectives "not" and "if–then."

Section A2.2. "Not" and "If-Then"

The statement "not **A**," written $\sim \mathbf{A}$, is true whenever **A** is false. For example, the statement

Gene is not tall.

is true provided the statement "Gene is tall" is false. The truth table for $\sim \mathbf{A}$ is as follows

A	**∼ A**
T	F
F	T

Although "not" is a simple idea, it can be a powerful tool when used in proofs by contradiction. To prove that a statement **A** is true using proof by contradiction, we instead assume ∼ **A**. We then show that this hypothesis leads to a contradiction. Thus ∼ **A** must be false; according to the truth table, we see that the only remaining possibility is that **A** is true.

Greater understanding is obtained by combining connectives:

Example A2.6

Here is the truth table for ∼ (**A** ∨ **B**):

A	**B**	**A** ∨ **B**	∼ (**A** ∨ **B**)
T	T	T	F
T	F	T	F
F	T	T	F
F	F	F	T

□

Example A2.7

Now we look at the truth table for (∼ **A**) ∧ (∼ **B**):

A	**B**	∼ **A**	∼ **B**	(∼ **A**) ∧ (∼ **B**)
T	T	F	F	F
T	F	F	T	F
F	T	T	F	F
F	F	T	T	T

□

Notice that the statements ∼ (**A** ∨ **B**) and (∼ **A**) ∧ (∼ **B**) have the *same truth table* (look at the last column in each table). We call such pairs of statements *logically equivalent*.

The logical equivalence of ∼ (**A** ∨ **B**) with (∼ **A**) ∧ (∼ **B**) makes good intuitive sense: the statement **A** ∨ **B** fails if and only if **A** is false *and* **B** is false. Since in mathematics we cannot rely on our intuition to establish facts, it is important to have the truth table technique for establishing logical equivalence.

A statement of the form "If **A** then **B**" asserts that whenever **A** is true then **B** is also true. This assertion (or "promise") is tested when **A** is true, because it is then claimed that something else (namely, **B**) is true as well. *However*, when **A** is false then the statement "If **A** then **B**" *claims nothing.*

Using the symbols $\mathbf{A} \Rightarrow \mathbf{B}$ to denote "If $\mathbf{A}$ then $\mathbf{B}$," we obtain the following truth table:

A	B	A $\Rightarrow$ B
T	T	T
T	F	F
F	T	T
F	F	T

Notice that we use here an important principle of Aristotelian logic: every sensible statement is either true or false. There is no "in between" status. Thus when $\mathbf{A}$ is false then the statement $\mathbf{A} \Rightarrow \mathbf{B}$ is not tested. It therefore cannot be false. So it must be true. In fact the only way that $\mathbf{A} \Rightarrow \mathbf{B}$ can be false is if $\mathbf{A}$ is true and $\mathbf{B}$ is false.

Example A2.8

The statement $\mathbf{A} \Rightarrow \mathbf{B}$ is logically equivalent with $\sim (\mathbf{A} \wedge \sim \mathbf{B})$. For the truth table for the latter is

A	B	$\sim$ B	A $\wedge \sim$ B	$\sim (\mathbf{A} \wedge \sim \mathbf{B})$
T	T	F	F	T
T	F	T	T	F
F	T	F	F	T
F	F	T	F	T

which is the same as the truth table for $\mathbf{A} \Rightarrow \mathbf{B}$.   □

There are in fact infinitely many pairs of logically equivalent statements. But just a few of these equivalences are really important in practice—most others are built up from these few basic ones.

Example A2.9

The statement

$$\textbf{If} \quad x \textbf{ is negative then} \quad -5 \cdot x \textbf{ is positive.}$$

is true. For if $x < 0$ then $-5 \cdot x$ is indeed > 0; if $x \geq 0$ then the statement is unchallenged.   □

Example A2.10

The statement

$$\text{If } \{x > 0 \text{ and } x^2 < 0\} \text{ then } x \geq 10.$$

is true since the hypothesis "$x > 0$ and $x^2 < 0$" is never true.

□

Example A2.11

The statement

$$\text{If } x > 0 \text{ then } \{x^2 < 0 \text{ or } 2x < 0\}.$$

is false since the conclusion "$x^2 < 0$ *or* $2x < 0$" is false whenever the hypothesis $x > 0$ is true. □

Section A2.3. Contrapositive, Converse, and "Iff"

The statement

$$\text{If } \mathbf{A} \text{ then } \mathbf{B}. \qquad \text{or} \qquad \mathbf{A} \Rightarrow \mathbf{B}.$$

is the same as saying

$$\mathbf{A} \text{ suffices for } \mathbf{B}.$$

or as saying

$$\mathbf{A} \text{ only if } \mathbf{B}.$$

All these forms are encountered in practice, and you should think about them long enough to realize that they all say the same thing.

On the other hand,

$$\text{If } \mathbf{B} \text{ then } \mathbf{A}. \qquad \text{or} \qquad \mathbf{B} \Rightarrow \mathbf{A}.$$

is the same as saying

$$\mathbf{A} \text{ is necessary for } \mathbf{B}.$$

or as saying

$$\mathbf{A} \text{ if } \mathbf{B}.$$

We call the statement $\mathbf{B} \Rightarrow \mathbf{A}$ the *converse* of $\mathbf{A} \Rightarrow \mathbf{B}$.

Example A2.12

The converse of the statement

$$\text{If } x \text{ is a healthy horse then } x \text{ has four legs.}$$

is the statement

If x has four legs then x is a healthy horse.

Notice that these statements have very different meanings: the first statement is true while the second (its converse) is false. For example, my desk has four legs but it is not a healthy horse.

□

The statement

A if and only if B.

is a brief way of saying

If A then B.   *and*   **If B then A.**

We abbreviate **A if and only if B**   as   **A $\Leftrightarrow$ B**   or as   **A iff B**. Here is a truth table for **A $\Leftrightarrow$ B**.

A	**B**	**A $\Rightarrow$ B**	**B $\Rightarrow$ A**	**A $\Leftrightarrow$ B**
T	T	T	T	T
T	F	F	T	F
F	T	T	F	F
F	F	T	T	T

Notice that we can say that **A $\Leftrightarrow$ B** is true only when both **A $\Rightarrow$ B** and **B $\Rightarrow$ A** are true. An examination of the truth table reveals that **A $\Leftrightarrow$ B** is true precisely when **A** and **B** are either both true or both false. Thus **A $\Leftrightarrow$ B** means precisely that   **A** and   **B** are logically equivalent. One is true *when and only when* the other is true.

Example A2.13

The statement
$$x > 0 \Leftrightarrow 2x > 0$$

is true. For if $x > 0$ then $2x > 0$; and if $2x > 0$ then $x > 0$.   □

Example A2.14

The statement
$$x > 0 \Leftrightarrow x^2 > 0$$

is false. For $x > 0 \Rightarrow x^2 > 0$ is certainly true while $x^2 > 0 \Rightarrow x > 0$ is false ($(-3)^2 > 0$ but $-3 \not> 0$).   □

Example A2.15

The statement

$$\{\sim (\mathbf{A} \vee \mathbf{B})\} \Leftrightarrow \{(\sim \mathbf{A}) \wedge (\sim \mathbf{B})\} \qquad (A2.15.1)$$

is true because the truth table for $\sim(\mathbf{A} \vee \mathbf{B})$ and that for $(\sim \mathbf{A}) \wedge (\sim \mathbf{B})$ are the same (we noted this fact in the last section). Thus they are logically equivalent: one statement is true precisely when the other is. Another way to see the truth of (A2.15.1) is to examine the full truth table:

A	B	$\sim (\mathbf{A} \vee \mathbf{B})$	$(\sim \mathbf{A}) \wedge (\sim \mathbf{B})$	$\sim (\mathbf{A} \vee \mathbf{B}) \Leftrightarrow \{(\sim \mathbf{A}) \wedge (\sim \mathbf{B})\}$
T	T	F	F	T
T	F	F	F	T
F	T	F	F	T
F	F	T	T	T

□

Given an implication

$$\mathbf{A} \Rightarrow \mathbf{B},$$

the *contrapositive* statement is defined to be the implication

$$\sim \mathbf{B} \Rightarrow \sim \mathbf{A}.$$

The contrapositive is logically equivalent to the original implication, as we see by examining their truth tables:

A	B	$\mathbf{A} \Rightarrow \mathbf{B}$
T	T	T
T	F	F
F	T	T
F	F	T

and

A	B	$\sim \mathbf{A}$	$\sim \mathbf{B}$	$(\sim \mathbf{B}) \Rightarrow (\sim \mathbf{A})$
T	T	F	F	T
T	F	F	T	F
F	T	T	F	T
F	F	T	T	T

Example A2.16

The statement

If it is raining, then it is cloudy.

has, as its contrapositive, the statement

If there are no clouds, then it is not raining.

A moment's thought convinces us that these two statements say the same thing: if there are no clouds, then it could not be raining; for the presence of rain implies the presence of clouds.

◻

Example A2.17

The statement

If X is a healthy horse then X has four legs.

has, as its contrapositive, the statement

If X does not have four legs then X is not a healthy horse.

A moment's thought reveals that these two statements say precisely the same thing. They are logically equivalent.

The main point to keep in mind is that, given an implication $\mathbf{A} \Rightarrow \mathbf{B}$, its *converse* $\mathbf{B} \Rightarrow \mathbf{A}$ and its *contrapositive* $(\sim \mathbf{B}) \Rightarrow (\sim \mathbf{A})$ are two different statements. The converse is distinct from, and *logically independent from*, the original statement. The contrapositive is distinct from, but *logically equivalent to*, the original statement.

Section A2.4. Quantifiers

The mathematical statements that we will encounter in practice will use the *connectives* "and," "or," "not," "if–then," and "iff." They will also use *quantifiers*. The two basic quantifiers are "for all" and "there exists."

Example A2.18

Consider the statement

All automobiles have wheels.

This statement makes an assertion about *all* automobiles. It is true, just because every automobile does have wheels.

Compare this statement with the next one:

There exists a woman who is blonde.

This statement is of a different nature. It does not claim that all women have blonde hair—merely that there exists *at least one* woman who does. Since that is true, the statement is true. □

Example A2.19

Consider the statement

All positive real numbers are integers.

This sentence asserts that something is true for all positive real numbers. It is indeed true for *some* positive real numbers, such as 1 and 2 and 193. However, it is false for at least one positive number (such as π), so the entire statement is false.

Here is a more interesting example:

The square of any real number is positive.

This assertion is *almost* true—the only exception is the real number 0: we see that $0^2 = 0$ is not positive. But it only takes one exception to falsify a "for all" statement. So the assertion is false. □

Example A2.20

Look at the statement

There exists a real number which is greater than 4.

In fact there are lots of real numbers which are greater than 4; some examples are 7, 8π, and 97/3. Since there is *at least one* number satisfying the assertion, the assertion is true.

A somewhat different example is the sentence

There exists a real number which satisfies the equation $$x^3 + x^2 + x + 1 = 0.$$

There is in fact only one real number which satisfies the equation, and that is $x = -1$. Yet that information is sufficient to make the statement true. □

We often use the symbol $\forall$ to denote "for all" and the symbol $\exists$ to denote "there exists." The assertion

$$\forall x, \ x + 1 < x$$

claims that, for every x, the number $x+1$ is less than x. If we take our universe to be the standard real number system, this statement is false (for example, $5+1$ is not less than 5). The assertion

$$\exists x, \ x^2 = x$$

claims that there is a number whose square equals itself. If we take our universe to be the real numbers, then the assertion is satisfied by $x = 0$ and by $x = 1$. Therefore the assertion is true.

Quite often we will encounter $\forall$ and $\exists$ used together. The following examples are typical:

Example A2.21

The statement

$$\forall x \ \exists y, \ y > x$$

claims that for any number x there is a number y which is greater than it. In the realm of the real numbers this is true. In fact $y = x + 1$ will always do the trick.

The statement

$$\exists x \ \forall y, \ y > x$$

has quite a different meaning from the first one. It claims that there is an x which is less than *every* y. This is absurd. For instance, x is *not* less than $y = x - 1$. ☐

Example A2.22

The statement

$$\forall x \ \forall y, \ x^2 + y^2 \geq 0$$

is true in the realm of the real numbers: it claims that the sum of two squares is always greater than or equal to zero.

The statement

$$\exists x \ \exists y, \ x + 2y = 7$$

is true in the realm of the real numbers: it claims that there exist x and y such that $x + 2y = 7$. The numbers $x = 3, y = 2$ will do the job (although there are many other choices that work as well). ☐

We conclude by noting that $\forall$ and $\exists$ are closely related. The statements

$$\forall x, \ B(x) \qquad \text{and} \qquad \sim \exists x, \ \sim B(x)$$

are logically equivalent. The first asserts that the statement $B(x)$ is true for all values of x. The second asserts that there exists no value of x for which $B(x)$ fails, which is the same thing.

Likewise, the statements

$$\exists x, B(x) \qquad \text{and} \qquad \sim \forall x, \sim B(x)$$

are logically equivalent. The first asserts that there is some x for which $B(x)$ is true. The second claims that it is not the case that $B(x)$ fails for every x, which is the same thing.

Remark A2.23

Most of the statements that we encounter in mathematics are formulated using "for all" and "there exists." For example,

Through every point P not on a line ℓ there is a line parallel to ℓ.

Each continuous function on a closed, bounded interval has an absolute maximum.

Each of these statements uses (implicitly) both a "for all" and a "there exists." A "for all" statement is like an *infinite conjunction*. The statement $\forall x, P(x)$ (when x is a natural number, let us say) says $P(1) \wedge P(2) \wedge P(3) \wedge \cdots$. A "there exists" statement is like an *infinite disjunction*. The statement $\exists x, Q(x)$ (when x is a natural number, let us say) says $Q(1) \vee Q(2) \vee Q(3) \vee \cdots$. Thus it is neither practical nor sensible to endeavor to verify statements such as these using truth tables. This is one of the chief reasons that we learn to produce mathematical proofs. One of the main themes of the present text is to gain new insights and to establish facts about the real number system using mathematical proofs.

Section A2.5. Set Theory and Venn Diagrams

The two most basic objects in all of mathematics are sets and functions. In this section we discuss the first of these two concepts.

A *set* is a collection of objects. For example, "the set of all blue shirts" and "the set of all lonely whales" are two examples of sets. In mathematics, we often write sets with the following "set-builder" notation:

$$\{x : x + 5 > 0\}.$$

This is read "the set of all x such that $x+5$ is greater than 0." The universe from which x is chosen (for us this will usually be the real numbers) is understood from context, though sometimes we may be more explicit and write

$$\{x \in \mathbb{R} : x + 5 > 0\}.$$

Here $\in$ is a symbol that means "is an element of."

Notice that the role of x in the set-builder notation is as a *dummy variable*; the set we have just described could also be written as

$$\{s : s + 5 > 0\}$$

or

$$\{\alpha \in \mathbb{R} : \alpha + 5 > 0\}.$$

To repeat, the symbol $\in$ is used to express membership in a set; for example, the statement

$$4 \in \{x : x > 0\}$$

says that 4 is a member of (or *an element of*) the set of all numbers x which are greater than 0. In other words, 4 is a positive number.

If A and B are sets, then the statement

$$A \subset B$$

is read "A is a subset of B." It means that each element of A is also an element of B (but not vice versa!). In other words $\boldsymbol{x \in A \Rightarrow x \in B}$.

Example A2.24

Let

$$A = \{x \in \mathbb{R} : \exists y \text{ such that } x = y^2\}$$

and

$$B = \{t \in \mathbb{R} : t + 3 > -5\}.$$

Then $A \subset B$. Why? The set A consists of those numbers that are squares— that is, A is just the nonnegative real numbers. The set B contains all numbers which are greater than -8. Since every nonnegative number (element of A) is also greater than -8 (element of B), it is correct to say that $A \subset B$.

However, it is not correct to say that $B \subset A$, because -2 is an element of B but is not an element of A.   □

We write $A = B$ to indicate that both $A \subset B$ and $B \subset A$. In these circumstances we say that the two sets are equal: every element of A is an element of B and every element of B is an element of A.

We use a slash through the symbols $\in$ or $\subset$ to indicate negation:

$$-4 \notin \{x : x \geq -2\}$$

and

$$\{x : x = x^2\} \not\subset \{y : y > 1/2\}.$$

It is often useful to combine sets. The set $A \cup B$, called the *union* of A and B, is the set consisting of all objects which are either elements of A *or* elements of B (or both). The set $A \cap B$, called the *intersection* of A and B, is the set consisting of all objects which are elements of *both* A and B.

Example A2.25

Let
$$A = \{x : -4 < x \le 3\} \quad , \quad B = \{x : -1 \le x < 7\},$$
$$C = \{x : -9 \le x \le 12\}.$$

Then
$$A \cup B = \{x : -4 < x < 7\} \quad A \cap B = \{x : -1 \le x \le 3\},$$
$$B \cup C = \{x : -9 \le x \le 12\} \quad , \quad B \cap C = \{x : -1 \le x < 7\}.$$

Notice that $B \cup C = C$ and $B \cap C = B$ because $B \subset C$. $\qquad \square$

Example A2.26

Let
$$A = \{\alpha \in \mathbb{Z} : \alpha \ge 9\}$$
$$B = \{\beta \in \mathbb{R} : -4 < \beta \le 24\},$$
$$C = \{\gamma \in \mathbb{R} : 13 < \gamma \le 30\}.$$

Then
$$(A \cap B) \cap C = \{x \in \mathbb{Z} : 9 \le x \le 24\} \cap C = \{t \in \mathbb{Z} : 13 < t \le 24\}.$$

Also
$$A \cap (B \cup C) = A \cap \{x \in \mathbb{R} : -4 < x \le 30\} = \{y \in \mathbb{Z} : 9 \le x \le 30\}.$$

Try your hand at calculating $A \cup (B \cup C)$. $\qquad \square$

The symbol $\emptyset$ is used to denote the set with no elements. We call this set the *empty set*. For instance,
$$A = \{x \in \mathbb{R} : x^2 < 0\}$$

is a perfectly good set. However, there are no real numbers which satisfy the given condition. Thus A is empty, and we write $A = \emptyset$.
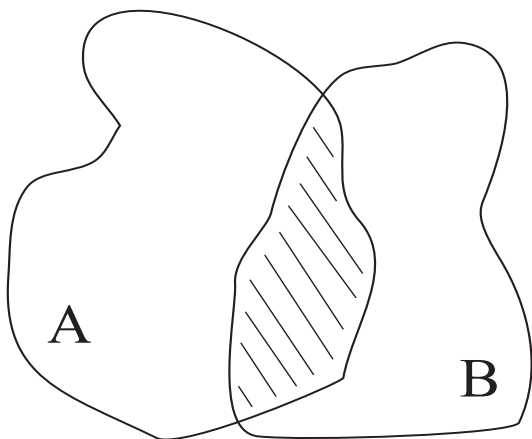
Figure A2.1: The intersection of two sets.

Example A2.27

Let
$$A = \{x : x > 8\} \quad \text{and} \quad B = \{x : x^2 < 4\}.$$
Then $A \cup B = \{x : x > 8 \text{ or } -2 < x < 2\}$ while $A \cap B = \emptyset$. $\qquad\square$

We sometimes use a *Venn diagram* to aid our understanding of set-theoretic relationships. In a Venn diagram, a set is represented as a domain in the plane. The intersection $A \cap B$ of two sets A and B is the region common to the two domains—see Figure A2.1.

Now let A, B, and C be three sets. The Venn diagram in Figure A2.2 makes it easy to see that $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$.

If A and B are sets then $A \setminus B$ denotes those elements which are *in A* but *not in B*. This operation is sometimes called *subtraction of sets* or *set-theoretic difference*.

Example A2.28

Let
$$A = \{x : 4 < x\}$$
and
$$B = \{x : 6 \le x \le 8\}.$$
Then
$$A \setminus B = \{x : 4 < x < 6\} \cup \{x : 8 < x\}$$
while
$$B \setminus A = \emptyset.$$
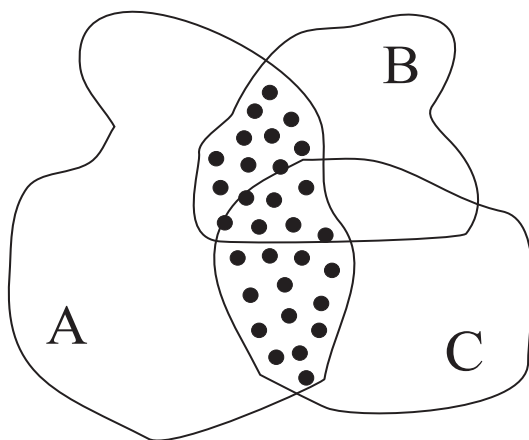
Figure A2.2: $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$.

Notice that $A \setminus A = \emptyset$; this fact is true for any set. □

Example A2.29

Let
$$S = \{x : 5 \le x\}$$
and
$$T = \{x : 4 < x < 6\} \, .$$
Then
$$S \setminus T = \{x : 6 \le x\} \qquad \text{and} \qquad T \setminus S = \{x : 4 < x < 5\} \, .$$

The Venn diagram in Figure A2.3 illustrates the fact that

$$A \setminus (B \cup C) = (A \setminus B) \cap (A \setminus C)$$

A Venn diagram is not a proper substitute for a rigorous mathematical proof. However, it can go a long way toward guiding our intuition.

We conclude this section by mentioning a useful set-theoretic operation and an application. Suppose that we are studying subsets of a fixed set X. We sometimes call X the "universal set." If $S \subset X$ then we use the notation ${}^c S$ to denote the set $X \setminus S$ or $\{x \in X : x \notin S\}$. The set ${}^c S$ is called *the complement of S* (in the set X).
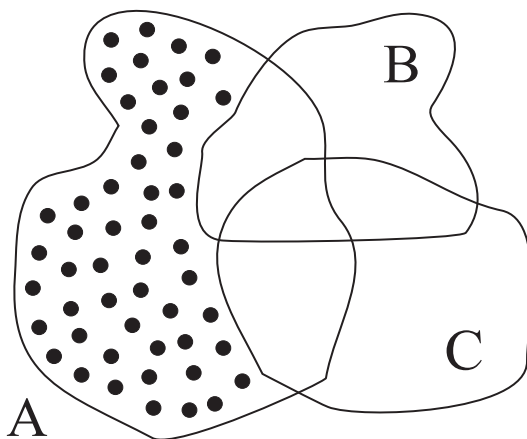
Figure A2.3: $A \setminus (B \cup C) = (A \setminus B) \cap (A \setminus C)$.

Example A2.30

When we study real analysis, most sets that we consider are subsets of the real line $\mathbb{R}$. If $S = \{x \in \mathbb{R} : 0 \le x \le 5\}$ then ${}^cS = \{x \in \mathbb{R} : x < 0\} \cup \{x \in \mathbb{R} : x > 5\}$. If T is the set of rational numbers then ${}^cT$ is the set of irrational numbers. □

If A, B are sets then it is straightforward to verify that ${}^c(A \cup B) = {}^cA \cap {}^cB$ and ${}^c(A \cap B) = {}^cA \cup {}^cB$. These are known as *de Morgan's laws.* Let us prove the first of these.

If $x \in {}^c(A \cup B)$ then x is not an element of $A \cup B$. Hence x is not an element of A and x is not an element of B. So $x \in {}^cA$ and $x \in {}^cB$. Therefore $x \in {}^cA \cap {}^cB$. That shows that ${}^c(A \cup B) \subset {}^cA \cap {}^cB$. For the reverse direction, assume that $x \in {}^cA \cap {}^cB$. Then $x \in {}^cA$ and $x \in {}^cB$. As a result, $x \notin A$ and $x \notin B$. So $x \notin A \cup B$. So $x \in {}^c(A \cup B)$. This shows that ${}^cA \cap {}^cB \subset {}^c(A \cup B)$.

The two inclusions that we have proved establish that ${}^c(A \cup B) = {}^cA \cap {}^cB$.

Section A2.6. Relations and Functions

In more elementary mathematics courses we learn that a "relation" is a rule for associating elements of two sets; and a "function" is a rule that associates to each element of one set a unique element of another set. The trouble with these definitions is that they are imprecise. For example, suppose we define the function $f(x)$ to be identically equal to 1 if there is life as we know it on Mars and to be identically equal to 0 if there is no life as we know it on Mars. Is this a good definition? It certainly is not a very practical one!

Value of Yen Against Dollar
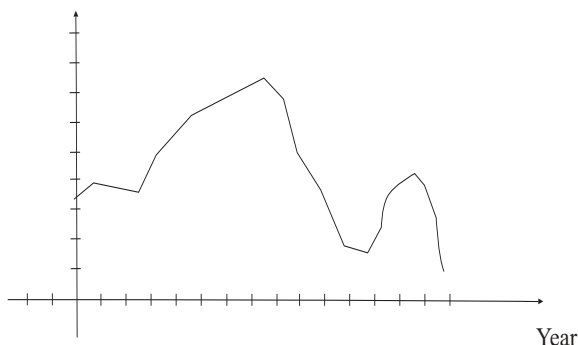
Figure A2.4: Value of the Yen against the Dollar.

More important is the fact that using the word "rule" suggests that functions are given by formulas. Indeed, some functions are; but most are not. Look at any graph in the newspaper—of unemployment, or the value of the Japanese Yen (Figure A2.4), or the Gross National Product. The graphs represent values of these parameters as a function of time. And it is clear that the functions are not given by elementary formulas.

To summarize, we need a notion of function, and of relation, which is precise and flexible and which does not tie us to formulas. We begin with relations, and then specialize down to functions.

Definition A2.31

Let A and B be sets. A *relation* $\mathcal{R}$ on A and B is a collection of ordered pairs (a, b) such that $a \in A$ and $b \in B$. (Notice that we did not say "*the* collection of all ordered pairs"—that is, a relation consists of some of the ordered pairs, but not necessarily all of them.) If a is related to b then we sometimes write $a\mathcal{R}b$ or $(a, b) \in \mathcal{R}$.

Example A2.32

Let A be the real numbers and B the integers. The set

$$\mathcal{R} = \{(\pi, 2), (3.4, -2), (\sqrt{2}, 94), (\pi, 50), (2 + \sqrt{17}, -2)\}$$

is a relation on A and B. It associates certain elements of A to certain elements of B. Observe that repetitions are allowed: $\pi \in A$ is associated to both 2 and 50 in B; also $-2 \in B$ is associated to both 3.4 and $2 + \sqrt{17}$ in A. This relation is not given by any formula or rule.

Now let

$$A = \{3, 17, 28, 42\} \quad \text{and} \quad B = \{10, 20, 30, 40\}.$$

Then

$$\mathcal{R} = \{(3, 10), (3, 20), (3, 30), (3, 40), (17, 20), (17, 30),$$

$$(17, 40), (28, 30), (28, 40)\}$$

is a relation on A and B. In fact $a \in A$ is related to $b \in B$ precisely when $a < b$. This second relation *is* given by a rule. □

Example A2.33

Let

$$A = B = \{\text{meter, pound, foot, ton, yard, ounce}\}.$$

Then

$$\mathcal{R} = \{(\text{foot,meter}), (\text{foot,yard}), (\text{meter,yard}), (\text{pound,ton}),$$
$$(\text{pound,ounce}),(\text{ton,ounce}), (\text{meter,foot}), (\text{yard,foot}),$$
$$(\text{yard,meter}),(\text{ton,pound}),(\text{ounce,pound}), (\text{ounce,ton})\}$$

is a relation on A and B. In fact two words are related by $\mathcal{R}$ if and only if they measure the same thing: foot, meter, and yard measure length while pound, ton, and ounce measure weight.

Notice that the pairs in $\mathcal{R}$, and in any relation, are *ordered* pairs: the pair (foot,yard) is different from the pair (yard,foot). □

Example A2.34

Let

$$A = \{25, 37, 428, 695\} \quad \text{and} \quad B = \{14, 7, 234, 999\}$$

Then

$$\mathcal{R} = \{(25, 234), (37, 7), (37, 234), (428, 14), (428, 234), (695, 999)\}$$

is a relation on A and B. In fact two elements are related by $\mathcal{R}$ if and only if they have at least one digit in common. □

Definition A2.35

A relation $\mathcal{R}$ on a set A is said to be an *equivalence relation* if it has these three properties:

Reflexive: For any $a \in A$ it holds that $a\mathcal{R}a$.

Symmetric: If $a\mathcal{R}b$ then $b\mathcal{R}a$.

Transitive: If $a\mathcal{R}b$ and $b\mathcal{R}c$, then $a\mathcal{R}c$.

It can be proved that, if $\mathcal{R}$ is an equivalence relation, then it partitions A into pairwise disjoint equivalence classes. That is to say, if $x \in A$, then let

$$E_a = \{a \in A : a\mathcal{R}x\}.$$

We call E_a the *equivalence class* of a. Then it is the case that if $E_a \cap E_b \neq \emptyset$, then $E_a = E_b$. So the union of the E_a is all of A, and the E_a are pairwise disjoint. For all the details of the theory of equivalence classes, consult [KRA1].

A function is a special type of relation, as we shall now learn.

Definition A2.36

Let A and B be sets. A *function* from A to B is a relation $\mathcal{R}$ on A and B such that for each $a \in A$ there is one and only one pair $(a, b) \in \mathcal{R}$. We call A the *domain* of the function and we call B the *range*.[1]

Example A2.37

Let

$$A = \{1, 2, 3, 4\} \quad \text{and} \quad B = \{\alpha, \beta, \gamma, \delta\}.$$

Then

$$\mathcal{R} = \{(1, \gamma), (2, \delta), (3, \gamma), (4, \alpha)\}$$

is a function from A to B. Notice that there is precisely one pair in $\mathcal{R}$ for each element of A. However, notice that repetition of elements of B is allowed. Notice also that there is no apparent "pattern" or "rule" that determines $\mathcal{R}$. Finally observe that not all the elements of B are used.

With the same sets A and B consider the relations

$$S = \{(1, \alpha), (2, \beta), (3, \gamma)\}$$

and

$$\mathcal{T} = \{(1, \alpha), (2, \beta), (3, \gamma), (4, \delta), (2, \gamma)\}.$$

Then S is not a function because it violates the rule that there be a pair for *each* element of A. Also $\mathcal{T}$ is not a function because it violates the rule that there be *just one* pair for each element of A. □

The relations and function described in the last example were so simple that you may be wondering what happened to the kinds of functions that we usually look at in mathematics. Now we consider some of those.

---

[1]Some textbooks use the word "codomain" instead of "range." We shall use only the word "range."

Example A2.38

Let $A = \mathbb{R}$ and $B = \mathbb{R}$, where $\mathbb{R}$ denotes the real numbers. The relation

$$\mathcal{R} = \{(x, \sin x) : x \in A\}$$

is a function from A to B. For each $a \in A = \mathbb{R}$ there is one and only one ordered pair with first element a.

Now let $S = \mathbb{R}$ and $T = \{x \in \mathbb{R} : -2 \leq x \leq 2\}$. Then

$$\mathcal{U} = \{(x, \sin x) : x \in A\}$$

is also a function from S to T. Technically speaking, it is a different function from $\mathcal{R}$ because it has a different range. However, this distinction often has no practical importance and we shall not mention the difference. It is frequently convenient to write functions like $\mathcal{R}$ or $\mathcal{U}$ as

$$\mathcal{R}(x) = \sin x$$

and

$$\mathcal{U}(x) = \sin x \,. \qquad \square$$

The last example suggests that we distinguish between the set B where a function takes its values and the set of values that the function *actually assumes*.

Definition A2.39

Let A and B be sets and let f be a function from A to B. Define the *image* of f to be
$$\text{Image } f = \{b \in B : \exists a \in A \text{ such that } f(a) = b\} \,.$$

The set Image f is a subset of the range B. In general the image *will not* equal the range.

Example A2.40

Both the functions $\mathcal{R}$ and $\mathcal{U}$ from the last example have the set $\{x \in \mathbb{R} : -1 \leq x \leq 1\}$ as image. In neither instance does the image equal the range. $\square$

If a function f has domain A and range B and if S is a subset of A then we define
$$f(S) = \{b \in B : b = f(s) \text{ for some } s \in S\} \,.$$

The set $f(A)$ equals the image of f.

Example A2.41

Let $A = \mathbb{R}$ and $B = \{0, 1\}$. Consider the function

$$f = \{(x, y) : y = 0 \text{ if } x \text{ is rational and}$$

$$y = 1 \text{ if } x \text{ is irrational}\}.$$

The function f is called the *Dirichlet function* (P. G. Lejeune-Dirichlet, 1805–1859). It is given by a rule, but not by a formula.

Notice that $f(\mathbb{Q}) = \{0\}$ and $f(\mathbb{R}) = \{0, 1\}$. □

Definition A2.42

Let A and B be sets and f a function from A to B.

We say that f is *one-to-one* if whenever $(a_1, b) \in f$ and $(a_2, b) \in f$ then $a_1 = a_2$.

We say that f is *onto* if whenever $b \in B$ then there exists an $a \in A$ such that $(a, b) \in f$.

Example A2.43

Let $A = \mathbb{R}$ and $B = \mathbb{R}$. Consider the functions

$$f(x) = 2x + 5 \quad , \quad g(x) = \arctan x$$

$$h(x) = \sin x \quad , \quad j(x) = 2x^3 + 9x^2 + 12x + 4.$$

Then f is both one-to-one and onto, g is one-to-one but not onto, j is onto but not one-to-one, and h is neither.

Refer to to convince yourself of these assertions. □

When a function f is both one-to-one and onto then it is called a *bijection* of its domain to its range. Sometimes we call such a function a *set-theoretic isomorphism*. In the last example, the function f is a bijection of $\mathbb{R}$ to $\mathbb{R}$.

If f and g are functions, and if the image of g is contained in the domain of f, then we define the *composition* $f \circ g$ to be

$$\{(a, c) : \exists b \text{ such that } g(a) = b \text{ and } f(b) = c\}.$$

This may be written more simply as

$$f \circ g(a) = f(g(a)) = f(b) = c.$$

Let f have domain A and range B. Assume for simplicity that the image of f is all of B. If there exists a function g with domain B and range A such that

$$f \circ g(b) = b \quad \forall b \in B$$
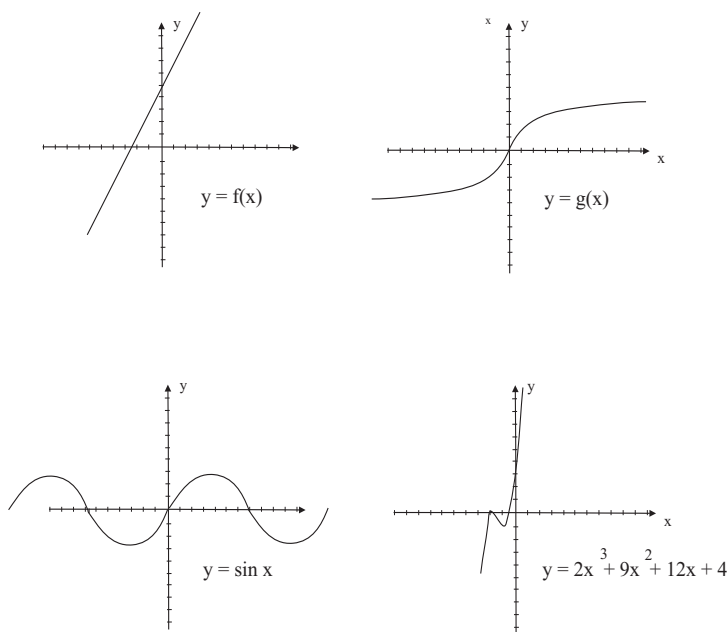
Figure A2.5: One-to-one and onto functions.

and

$$g \circ f(a) = a \quad \forall a \in A \,,$$

then g is called the *inverse* of f.

Clearly, if the function f is to have an inverse, then f must be one-to-one. For if $f(a) = f(a') = b$ then it cannot be that both $g(b) = a$ and $g(b) = a'$. Also f must be onto. For if some $b \in B$ is not in the image of f then it cannot hold that $f \circ g(b) = b$. It turns out that these two conditions are also sufficient for the function f to have an inverse: If f has domain A and range B and if f is both one-to-one and onto, then f has an inverse.

Example A2.44

Define a function f, with domain $\mathbb{R}$ and range $\{x \in \mathbb{R} : x \geq 0\}$, by the formula $f(x) = x^2$. Then f is onto but is not one-to-one (because $f(-1) = f(1)$), hence it cannot have an inverse. This is another way of saying that a positive real number has two square roots—not one.

However, the function g, with domain $\{x \in \mathbb{R} : x \geq 0\}$ and range $\{x \in \mathbb{R} : x \geq 0\}$, given by the formula $g(x) = x^2$, *does* have an inverse. In fact the inverse function is $h(x) = +\sqrt{x}$.

The function $k(x) = x^3$, with domain $\mathbb{R}$ and range $\mathbb{R}$, is both one-to-one and onto. It therefore has an inverse: the function $m(x) = x^{1/3}$ satisfies $k \circ m(x) = x$,

and $m \circ k(x) = x$ for all x. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad$ $\square$

Section A2.7. Countable and Uncountable Sets

One of the most profound ideas of modern mathematics is Georg Cantor's theory of the infinite (George Cantor, 1845–1918). Cantor's insight was that infinite sets can be compared by size, just as finite sets can. For instance, we think of the number 2 as *less* than the number 3; so a set with two elements is "smaller" than a set with three elements. We would like to have a similar notion of comparison for infinite sets. In this section we will present Cantor's ideas; we will also give precise definitions of the terms "finite" and "infinite."

Definition A2.45

Let A and B be sets. We say that A and B have the *same cardinality* if there is a function f from A to B which is both one-to-one and onto (that is, f is a bijection from A to B). We write $\mathrm{card}(A) = \mathrm{card}(B)$. Some books write $|A| = |B|$.

Example A2.46

Let $A = \{1, 2, 3, 4, 5\}, B = \{\alpha, \beta, \gamma, \delta, \epsilon\}, C = \{a, b, c, d, e, f\}$. Then A and B have the same cardinality because the function

$$f = \{(1, \alpha), (2, \beta), (3, \gamma), (4, \delta), (5, \epsilon)\}$$

is a bijection of A to B. This function is not the *only* bijection of A to B (can you find another?), but we are only required to produce one.

On the other hand, A and C do not have the same cardinality; neither do B and C. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad$ $\square$

Notice that if $\mathrm{card}(A) = \mathrm{card}(B)$ via a function f_1 and $\mathrm{card}(B) = \mathrm{card}(C)$ via a function f_2 then $\mathrm{card}(A) = \mathrm{card}(C)$ via the function $f_2 \circ f_1$.

Example A2.47

Let A and B be sets. If there is a one-to-one function from A to B but no bijection between A and B then we will write

$$\mathrm{card}(A) < \mathrm{card}(B).$$

This notation is read "A has smaller cardinality than B."

We use the notation

$$\mathrm{card}(A) \leq \mathrm{card}(B)$$

to mean that either $\operatorname{card}(A) < \operatorname{card}(B)$ or $\operatorname{card}(A) = \operatorname{card}(B)$.

Example A2.48

An extremely simple example of this last concept is given by $A = \{1, 2, 3\}$ and $B = \{a, b, c, d, e\}$. Then the function

$$
\begin{array}{rcl}
f : A & \to & B \\
1 & \mapsto & a \\
2 & \mapsto & b \\
3 & \mapsto & c
\end{array}
$$

is a one-to-one function from A to B. But there is no one-to-one function from B to A. We write

$$\operatorname{card}(A) < \operatorname{card}(B).$$

We shall see more profound applications, involving infinite sets, in our later discussions.   □

Notice that $\operatorname{card}(A) \leq \operatorname{card}(B)$ and $\operatorname{card}(B) \leq \operatorname{card}(C)$ imply that $\operatorname{card}(A) \leq \operatorname{card}(C)$. Moreover, if $A \subset B$, then the inclusion map $i(a) = a$ is a one-to-one function of A into B; therefore $\operatorname{card}(A) \leq \operatorname{card}(B)$.

The next theorem gives a useful method for comparing the cardinality of two sets.

Theorem A2.49   (Schroeder–Bernstein)

Let A, B, be sets. If there is a one-to-one function $f : A \to B$ and a one-to-one function $g : B \to A$, then A and B have the same cardinality.

Proof: It is convenient to assume that A and B are disjoint; we may do so by replacing A by $\{(a, 0) : a \in A\}$ and B by $\{(b, 1) : b \in B\}$. Let D be the image of f and C be the image of g. Let us define a *chain* to be a sequence of elements of either A or B—that is, a function $\phi : \mathbb{N} \to (A \cup B)$—such that

- $\phi(1) \in B \setminus D$;

- If for some j we have $\phi(j) \in B$, then $\phi(j + 1) = g(\phi(j))$;

- If for some j we have $\phi(j) \in A$, then $\phi(j + 1) = f(\phi(j))$.

We see that a chain is a sequence of elements of $A \cup B$ such that the first element is in $B \setminus D$, the second in A, the third in B, and so on. Obviously each element of $B \setminus D$ occurs as the first element of at least one chain.

Define $\mathcal{S} = \{a \in A : a$ is some term of some chain$\}$. It is helpful to note that

$$\mathcal{S} = \{x : x \text{ can be written in the form}$$
$$g(f(g(\cdots g(y)\dots))) \text{ for some } y \in B \setminus D\}. \qquad \text{(A2.49.1)}$$

We set
$$k(x) = \begin{cases} f(x) & \text{if} \quad x \in A \setminus \mathcal{S} \\ g^{-1}(x) & \text{if} \quad x \in \mathcal{S} \end{cases}$$

Note that the second half of this definition makes sense because $\mathcal{S} \subseteq C$. Then $k : A \to B$. We shall show that in fact k is a bijection.

First notice that f and g^{-1} are one-to-one. This is not quite enough to show that k is one-to-one, but we now reason as follows: If $f(x_1) = g^{-1}(x_2)$ for some $x_1 \in A \setminus \mathcal{S}$ and some $x_2 \in \mathcal{S}$, then $x_2 = g(f(x_1))$. But, by (A2.49.1), the fact that $x_2 \in \mathcal{S}$ now implies that $x_1 \in S$. That is a contradiction. Hence k is one-to-one.

It remains to show that k is onto. Fix $b \in B$. We seek an $x \in A$ such that $k(x) = b$.

Case A: If $g(b) \in \mathcal{S}$, then $k(g(b)) \equiv g^{-1}(g(b)) = b$ hence the x that we seek is $g(b)$.

Case B: If $g(b) \notin \mathcal{S}$, then we claim that there is an $x \in A$ such that $f(x) = b$. Assume this claim for the moment.

Now the x that we found in the last paragraph must lie in $A \setminus \mathcal{S}$. For if not then x would be in some chain. Then $f(x)$ and $g(f(x)) = g(b)$ would also lie in that chain. Hence $g(b) \in \mathcal{S}$, and that is a contradiction. But $x \in A \setminus \mathcal{S}$ tells us that $k(x) = f(x) = b$. That completes the proof that k is onto. Hence k is a bijection.

To prove the claim in Case B, notice that if there is no x with $f(x) = b$, then $b \in B \setminus D$. Thus some chain would begin at b. So $g(b)$ would be a term of that chain. Hence $g(b) \in \mathcal{S}$ and that is a contradiction.

The proof of the Schroeder–Bernstein theorem is complete. □

Remark A2.50

Let us reiterate some of the earlier ideas in light of the Schroeder–Bernstein theorem. If A and B are sets and if there is a one-to-one function $f : A \to B$, then we know that $\text{card}(A) \leq \text{card}(B)$. If there is no one-to-one function $g : B \to A$, then we may write $\text{card}(A) < \text{card}(B)$. But if instead there *is* a one-to-one function $g : B \to A$, then $\text{card}(B) \leq \text{card}(A)$ and the Schroeder–Bernstein theorem guarantees therefore that $\text{card}(A) = \text{card}(B)$.

Now it is time to look at some specific examples.

Example A2.51

Let E be the set of all even integers and O the set of all odd integers. Then

$$\text{card}(E) = \text{card}(O).$$

Indeed, the function

$$f(j) = j + 1$$

is a bijection from E to O. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\square$

Example A2.52

Let E be the set of even integers. Then

$$\text{card}(E) = \text{card}(\mathbb{Z}).$$

The function

$$g(j) = j/2$$

is a bijection from E to $\mathbb{Z}$. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\square$

This last example is a bit surprising, for it shows that the set $\mathbb{Z}$ can be put in one-to-one correspondence with a proper subset E of itself. In other words, we are saying that the integers $\mathbb{Z}$ "have the same number of elements" as a proper subset of $\mathbb{Z}$. Such a phenomenon cannot occur with finite sets.

Example A2.53

We have

$$\text{card}(\mathbb{Z}) = \text{card}(\mathbb{N}).$$

We define the function f from $\mathbb{Z}$ to $\mathbb{N}$ as follows:

- $f(j) = -(2j + 1)$ if j is negative

- $f(j) = 2j + 2$ if j is positive or zero

The values that f takes on the negative numbers are $1, 3, 5, \ldots$, on the positive numbers are $4, 6, 8, \ldots$, and $f(0) = 2$. Thus f is one-to-one and onto. $\qquad\square$

Definition A2.54

If a set A has the same cardinality as $\mathbb{N}$ then we say that A is *countable*.

By putting together the preceding examples, we see that the set of even integers, the set of odd integers, and the set of all integers are examples of countable sets.

Example A2.55

The set of all ordered pairs of positive integers

$$S = \{(j, k) : j, k \in \mathbb{N}\}$$

is countable.

To see this we will use the Schroeder–Bernstein theorem. The function

$$f(j) = (j, 1)$$

is a one-to-one function from $\mathbb{N}$ to S. Also the function $g(j, k) = 2^j \cdot 3^k$ is a one-to-one function from S to $\mathbb{N}$. By the Schroeder–Bernstein theorem, S and $\mathbb{N}$ have the same cardinality; hence S is countable. □

Remark A2.56

You may check for yourself that the function $F(j, k) = 2^{j-1} \cdot (2k - 1)$ is an explicit bijection from S to $\mathbb{N}$.

Since there is a bijection of the set of *all* integers with the set $\mathbb{N}$, it follows from the last example that the set of all pairs of integers (positive *and* negative) is countable.

Notice that the word "countable" is a good descriptive word: if S is a countable set then we can think of S as having a first element (the one corresponding to $1 \in \mathbb{N}$), a second element (the one corresponding to $2 \in \mathbb{N}$), and so forth. Thus we write $S = \{s(1), s(2), \dots\} = \{s_1, s_2, \dots\}$.

Definition A2.57

A nonempty set S is called *finite* if there is a bijection of S with a set of the form $\{1, 2, \dots, n\}$ for some positive integer n. If no such bijection exists, then the set is called *infinite*.

An important property of the natural numbers $\mathbb{N}$ is that any subset $S \subset \mathbb{N}$ has a least element. This is known as the Well Ordering Principle, and is studied in a course on logic. In the present text we take the properties of the natural numbers as given. We use some of these properties in the next proposition.

Proposition A2.58

If S is a countable set and R is a subset of S then either R is empty or R is finite or R is countable.

Proof: Assume that R is not empty.

Write $S = \{s_1, s_2, \ldots\}$. Let j_1 be the least positive integer such that $s_{j_1} \in R$. Let j_2 be the least integer following j_1 such that $s_{j_2} \in R$. Continue in this fashion. If the process terminates at the nth step, then R is finite and has n elements.

If the process does not terminate, then we obtain an enumeration of the elements of R:

$$1 \longleftrightarrow s_{j_1}$$
$$2 \longleftrightarrow s_{j_2}$$
$$\ldots$$

etc.

All elements of R are enumerated in this fashion since $j_\ell \geq \ell$. Therefore R is countable. □

A set is called *denumerable* if it is either empty, finite, or countable. Notice that the word "denumerable" is not the same as "countable." In fact "countable" is just one instance of denumerable.

The set $\mathbb{Q}$ of all rational numbers consists of all expressions

$$\frac{a}{b},$$

where a and b are integers and $b \neq 0$. Thus $\mathbb{Q}$ can be identified with the set of all ordered pairs (a, b) of integers with $b \neq 0$. After discarding duplicates, such as $\frac{2}{4} = \frac{1}{2}$, and using Example A2.55 and Proposition A2.58, we find that the set $\mathbb{Q}$ is countable.

Theorem A2.59

Let S_1, S_2 be countable sets. Set $\mathcal{S} = S_1 \cup S_2$. Then $\mathcal{S}$ is countable.

Proof: Let us write

$$S_1 = \{s_1^1, s_2^1, \ldots\}$$
$$S_2 = \{s_1^2, s_2^2, \ldots\}.$$

If $S_1 \cap S_2 = \emptyset$ then the function

$$s_j^k \mapsto (j, k)$$

is a bijection of $\mathcal{S}$ with a subset of $\{(j,k) : j, k \in \mathbb{N}\}$. We proved earlier (Example A2.55) that the set of ordered pairs of elements of $\mathbb{N}$ is countable. By Proposition A2.58, $\mathcal{S}$ is countable as well.

If there exist elements which are common to S_1, S_2 then discard any duplicates. The same argument (use the preceding proposition) shows that $\mathcal{S}$ is countable. $\qquad\square$

Theorem A2.60

If S and T are each countable sets then so is

$$S \times T \equiv \{(s,t) : s \in S, t \in T\}\,.$$

Proof: Since S is countable there is a bijection f from S to $\mathbb{N}$. Likewise there is a bijection g from T to $\mathbb{N}$. Therefore the function

$$(f \times g)(s,t) = (f(s), g(t))$$

is a bijection of $S \times T$ with $\mathbb{N} \times \mathbb{N}$, the set of order pairs of positive integers. But we saw in Example A2.55 that the latter is a countable set. Hence so is $S \times T$. $\square$

Remark A2.61

We used the theorem as a vehicle for defining the concept of *set-theoretic product*: If A and B are sets then

$$A \times B \equiv \{(a,b) : a \in A, b \in B\}\,.$$

More generally, if $A_1, A_2, \ldots, A_k$ are sets then

$$A_1 \times A_2 \times \cdots \times A_k \equiv \{(a_1, a_2, \ldots, a_k) : a_j \in A_j \text{ for all } j = 1, \ldots, k\}\,.$$

Corollary A2.62

If $S_1, S_2, \ldots, S_k$ are each countable sets then so is the set

$$S_1 \times S_2 \times \cdots \times S_k = \{(s_1, \ldots, s_k) : s_1 \in S_1, \ldots, s_k \in S_k\}$$

consisting of all ordered $k-$tuples $(s_1, s_2, \ldots, s_k)$ with $s_j \in S_j$.

Proof: We may think of $S_1 \times S_2 \times S_3$ as $(S_1 \times S_2) \times S_3$. Since $S_1 \times S_2$ is countable (by the theorem) and S_3 is countable, then so is $(S_1 \times S_2) \times S_3 = S_1 \times S_2 \times S_3$ countable. Continuing in this fashion, we can see that any finite product of

countable sets is also a countable set.   □

　　We are accustomed to the union $A \cup B$ of two sets or, more generally, the union $A_1 \cup A_2 \cup \cdots \cup A_k$ of finitely many sets. But sometimes we wish to consider the union of infinitely many sets. Let $S_1, S_2, \ldots$ be countably many sets. We say that x is an element of

$$\bigcup_{j=1}^{\infty} S_j$$

if x is an element of at least one of the S_j.

Corollary A2.63

　　The countable union of countable sets is countable.

Proof: Let $A_1, A_2, \ldots$ each be countable sets. If the elements of A_j are enumerated as $\{a_k^j\}$ and if the sets A_j are pairwise disjoint then the correspondence

$$a_k^j \longleftrightarrow (j, k)$$

is one-to-one between the union of the sets A_j and the countable set $\mathbb{N} \times \mathbb{N}$. This proves the result when the sets A_j have no common element. If some of the A_j have elements in common then we discard duplicates in the union and use Proposition A2.58.   □

Proposition A2.64

　　The collection $\mathcal{P}$ of all polynomials with integer coefficients is countable.

Proof: Let $\mathcal{P}_k$ be the set of polynomials of degree k with integer coefficients. A polynomial p of degree k has the form

$$p(x) = p_0 + p_1 x + p_2 x^2 + \cdots + p_k x^k\,.$$

　　The identification

$$p(x) \longleftrightarrow (p_0, p_1, \ldots, p_k)$$

identifies the elements of $\mathcal{P}_k$ with the $(k+1)$-tuples of integers. By Corollary A2.62, it follows that $\mathcal{P}_k$ is countable. But then Corollary A2.63 implies that

$$\mathcal{P} = \bigcup_{j=0}^{\infty} \mathcal{P}_j$$

is countable.   □

Georg Cantor's remarkable discovery is that *not all infinite sets are count-able.* We next give an example of this phenomenon.

In what follows, a *sequence* on a set S is a function from $\mathbb{N}$ to S. We usually write such a sequence as $s(1), s(2), s(3), \ldots$ or as $s_1, s_2, s_3, \ldots$.

Example A2.65

There exists an infinite set which is not countable (we call such a set *un-countable*). Our example will be the set S of all sequences on the set $\{0, 1\}$. In other words, S is the set of all infinite sequences of 0s and 1s. To see that S is uncountable, assume the contrary. Then there is a first sequence

$$\mathcal{S}^1 = \{s_j^1\}_{j=1}^\infty,$$

a second sequence

$$\mathcal{S}^2 = \{s_j^2\}_{j=1}^\infty,$$

and so forth. This will be a complete enumeration of all the members of S. But now consider the sequence $\mathcal{T} = \{t_j\}_{j=1}^\infty$, which we construct as follows:

- If $s_1^1 = 0$ then set $t_1 = 1$; if $s_1^1 = 1$ then set $t_1 = 0$;

- If $s_2^2 = 0$ then set $t_2 = 1$; if $s_2^2 = 1$ then set $t_2 = 0$;

- If $s_3^3 = 0$ then set $t_3 = 1$; if $s_3^3 = 1$ then set $t_3 = 0$;

$$\ldots$$

- If $s_j^j = 0$ then set $t_j = 1$; if $s_j^j = 1$ then set $t_j = 0$;

etc.

Now the sequence $\mathcal{T}$ differs from the first sequence $\mathcal{S}^1$ in the first element: $t_1 \neq s_1^1$.

The sequence $\mathcal{T}$ differs from the second sequence $\mathcal{S}^2$ in the second element: $t_2 \neq s_2^2$.

And so on: the sequence $\mathcal{T}$ differs from the jth sequence $\mathcal{S}^j$ in the jth element: $t_j \neq s_j^j$. So the sequence $\mathcal{T}$ is not in the set S. But $\mathcal{T}$ is *supposed* to be in the set S because it is a sequence of 0s and 1s and all of these have been hypothesized to be enumerated.

This contradicts our assumption, so S must be uncountable.

$\square$

Example A2.66

Consider the set of all decimal representations of numbers—both terminating and non-terminating. Here a terminating decimal is one of the form

$$27.43926$$

while a non-terminating decimal is one of the form

$$3.14159265\ldots.$$

In the case of the non-terminating decimal, no repetition is implied; the decimal simply continues without cease.

Now the set of all those decimals containing only the digits 0 and 1 can be identified in a natural way with the set of sequences containing only 0 and 1 (just put commas between the digits). And we just saw that the set of such sequences is uncountable.

Since the set of all decimal numbers is an even bigger set, it must be uncountable also.

As you may know, the set of all decimals identifies with the set of all real numbers. We find then that the set $\mathbb{R}$ of all real numbers is uncountable. (Contrast this with the situation for the rationals.) In Chapter 1 we learned about how the real number system is constructed using just elementary set theory. □

It is an important result of set theory (due to Cantor) that, given any set S, the set of all subsets of S (called the *power set* of S) has strictly greater cardinality than the set S itself. As a simple example, let $S = \{a, b, c\}$. Then the set of all subsets of S is

$$\left\{\ \emptyset, \{a\}, \{b\}, \{c\}, \{a, b\}, \{a, c\}, \{b, c\}, \{a, b, c\}\ \right\}.$$

The set of all subsets has eight elements while the original set has just three.

Even more significant is the fact that if S is an infinite set then the set of all its subsets has greater cardinality than S itself. This is a famous theorem of Cantor. Thus there are infinite sets of arbitrarily large cardinality.

In some of the examples in this Appendix we constructed a bijection between a given set (such as $\mathbb{Z}$) and a proper subset of that set (such as E, the even integers). It follows from the definitions that this is possible only when the sets involved are infinite.

Appendix III: Review of Linear Algebra

Section A3.1. Linear Algebra Basics

When we first learn linear algebra, the subject is difficult because it is not usually presented in the context of applications. In the current text we see one of the most important applications of linear algebra: to provide a language in which to do analysis of several real variables. We now give a quick review of elementary linear algebra.

The principal properties of a vector space are that it have an additive structure and an operation of scalar multiplication. If $\mathbf{u} = (u_1, u_2, \ldots, u_k)$ and $\mathbf{v} = (v_1, v_2, \ldots, v_k)$ are elements of $\mathbb{R}^k$ and $a \in \mathbb{R}$ then define the operations of addition and scalar multiplication as follows:

$$\mathbf{u} + \mathbf{v} = (u_1 + v_1, u_2 + v_2, \ldots, u_k + v_k)$$

and

$$a \cdot \mathbf{u} = (au_1, au_2, \ldots, au_k).$$

Notice that the vector $\mathbf{0} = (0, 0, \ldots, 0)$ is the additive identity: $\mathbf{u} + \mathbf{0} = \mathbf{u}$ for any element $\mathbf{u} \in \mathbb{R}^k$. Also every element $\mathbf{u} = (u_1, u_2, \ldots, u_k) \in \mathbb{R}^k$ has an additive inverse $-\mathbf{u} = (-u_1, -u_2, \ldots, -u_k)$ that satisfies $\mathbf{u} + (-\mathbf{u}) = \mathbf{0}$.

Example A3.1

We have

$$(3, -2, 7) + (4, 1, -9) = (7, -1, -2)$$

and

$$5 \cdot (3, -2, 7, 14) = (15, -10, 35, 70). \qquad \Box$$

The first major idea in linear algebra is that of linear dependence:

Definition A3.2

A collection of elements $\mathbf{u}^1, \mathbf{u}^2, \ldots, \mathbf{u}^m \in \mathbb{R}^k$ is said to be *linearly dependent* if there exist constants $a_1, a_2, \ldots, a_m$, not all zero, such that

$$\sum_{j=1}^{m} a_j \mathbf{u}^j = 0.$$

Example A3.3

The vectors $\mathbf{u} = (1,3,4), \mathbf{v} = (2,-1,-3)$, and $\mathbf{w} = (5,1,-2)$ are linearly dependent because $1 \cdot \mathbf{u} + 2 \cdot \mathbf{v} - 1 \cdot \mathbf{w} = \mathbf{0}$.

However, the vectors $\mathbf{u}' = (1,0,0), \mathbf{v}' = (0,1,1)$, and $\mathbf{w}' = (1,0,1)$ are *not* linearly dependent since, if there were constants a, b, c such that

$$a\,\mathbf{u}' + b\,\mathbf{v}' + c\,\mathbf{w}' = \mathbf{0},$$

then

$$(a+c, b, b+c) = \mathbf{0}.$$

But this means that

$$\begin{aligned} a + c &= 0 \\ b &= 0 \\ b + c &= 0. \end{aligned}$$

We conclude that a, b, c must all be equal to zero. That is not allowed in the definition of linear dependence. $\square$

A collection of vectors that is not linearly dependent is called *linearly independent*. The vectors $\mathbf{u}', \mathbf{v}', \mathbf{w}'$ in the last example are linearly independent. Any set of k linearly independent vectors in $\mathbb{R}^k$ is called a **basis** for $\mathbb{R}^k$.

How do we recognize a basis? Notice that k vectors

$$\begin{aligned} \mathbf{u}^1 &= (u_1^1, u_2^1, \ldots, u_k^1) \\ \mathbf{u}^2 &= (u_1^2, v_2^2, \ldots, v_k^2) \\ &\quad \cdots \\ \mathbf{u}^k &= (u_1^k, u_2^k, \ldots, u_k^k) \end{aligned}$$

are linearly dependent if and only if there are numbers $a_1, a_2, \ldots, a_k$, not all zero, such that

$$a_1\,\mathbf{u}^1 + a_2\,\mathbf{u}^2 + \cdots + a_k\,\mathbf{u}^k = \mathbf{0}.$$

This in turn is true if and only if the system of equations

$$
\begin{aligned}
a_1 u_1^1 + a_2 u_1^2 + \cdots + a_k u_1^k &= 0 \\
a_1 u_2^1 + a_2 u_2^2 + \cdots + a_k u_2^k &= 0 \\
&\cdots \\
a_1 u_k^1 + a_2 u_k^2 + \cdots + a_k u_k^k &= 0
\end{aligned}
$$

has a nontrivial solution. But such a system has a nontrivial solution if and only if

$$
\det \begin{pmatrix}
u_1^1 & u_1^2 & \cdots & u_1^k \\
u_2^1 & u_2^2 & \cdots & u_2^k \\
& \cdots & \\
u_k^1 & u_k^2 & \cdots & u_k^k
\end{pmatrix} = 0.
$$

So a basis is a set of k vectors as above such that this determinant is *not* 0.

Bases are important because if $\mathbf{u}^1, \mathbf{u}^2, \ldots, \mathbf{u}^k$ form a basis then every element $\boldsymbol{x}$ of $\mathbb{R}^k$ can be expressed in one and only one way as

$$
\boldsymbol{x} = a_1 \mathbf{u}^1 + a_2 \mathbf{u}^2 + \cdots + a_k \mathbf{u}^k ,
$$

with $a_1, a_2, \ldots, a_k$ scalars. We call this a representation of $\boldsymbol{x}$ as a *linear combination* of $\mathbf{u}^1, \mathbf{u}^2, \ldots, \mathbf{u}^k$. To see that such a representation is always possible, and is unique, let $\boldsymbol{x} = (x_1, x_2, \ldots, x_k)$ be any element of $\mathbb{R}^k$. If $\mathbf{u}^1, \mathbf{u}^2, \ldots, \mathbf{u}^k$ form a basis then we wish to find $a_1, a_2, \ldots, a_k$ such that

$$
\boldsymbol{x} = a_1 \mathbf{u}^1 + a_2 \mathbf{u}^2 + \cdots + a_k \mathbf{u}^k .
$$

But, as above, this leads to the system of equations

$$
\begin{aligned}
a_1 u_1^1 + a_2 u_1^2 + \cdots + a_k u_1^k &= x_1 \\
a_1 u_2^1 + a_2 u_2^2 + \cdots + a_k u_2^k &= x_2 \\
&\cdots \\
a_1 u_k^1 + a_2 u_k^2 + \cdots + a_k^k u_k^k &= x_k .
\end{aligned}
\tag{A3.4}
$$

Now Cramer's Rule tells us that the unique solution of the system (A3.4) is given by

$$
a_1 = \frac{\det \begin{pmatrix}
x_1 & u_1^2 & \cdots & u_1^k \\
x_2 & u_2^2 & \cdots & u_2^k \\
& \cdots & \\
x_k & u_k^2 & \cdots & u_k^k
\end{pmatrix}}{\det \begin{pmatrix}
u_1^1 & u_1^2 & \cdots & u_1^k \\
u_2^1 & u_2^2 & \cdots & u_2^k \\
& \cdots & \\
u_k^1 & u_k^2 & \cdots & u_k^k
\end{pmatrix}} , \quad
a_2 = \frac{\det \begin{pmatrix}
u_1^1 & x_1 & \cdots & u_1^k \\
u_2^1 & x_2 & \cdots & u_2^k \\
& \cdots & \\
u_k^1 & x_k & \cdots & u_k^k
\end{pmatrix}}{\det \begin{pmatrix}
u_1^1 & u_1^2 & \cdots & u_1^k \\
u_2^1 & u_2^2 & \cdots & u_2^k \\
& \cdots & \\
u_k^1 & u_k^2 & \cdots & u_k^k
\end{pmatrix}} ,
$$

$$
\cdots
$$

$$\ldots, a_k = \frac{\det \begin{pmatrix} u_1^1 & u_1^2 & \cdots & x_1 \\ u_2^1 & u_2^2 & \cdots & x_2 \\ & \cdots & & \\ u_k^1 & u_k^2 & \cdots & x_k \end{pmatrix}}{\det \begin{pmatrix} u_1^1 & u_1^2 & \cdots & u_1^k \\ u_2^1 & u_2^2 & \cdots & u_2^k \\ & \cdots & & \\ u_k^1 & u_k^2 & \cdots & u_k^k \end{pmatrix}}.$$

Notice that the nonvanishing of the determinant in the denominator is crucial for this method to work.

In practice we will be given a basis $\mathbf{u}^1, \mathbf{u}^2, \ldots, \mathbf{u}^k$ for $\mathbb{R}^k$ and a vector $\boldsymbol{x}$ and we wish to express $\boldsymbol{x}$ as a linear combination of $\mathbf{u}^1, \mathbf{u}^2, \ldots, \mathbf{u}^k$. We may do so by solving a system of linear equations as above. A more elegant way to do this is to use the concept of the inverse of a matrix.

Definition A3.5

If
$$M = (m_{pq})_{\substack{p=1,\ldots,k \\ q=1,\ldots,\ell}}$$

is a $k \times \ell$ matrix (where k is the number of rows, ℓ the number of columns, and m_{pq} is the element in the pth row and qth column) and

$$N = (n_{rs})_{\substack{r=1,\ldots,\ell \\ s=1,\ldots,m}}$$

is an $\ell \times m$ matrix, then the *product* $M \cdot N$ is defined to be the matrix

$$T = (t_{uv})_{\substack{u=1,\ldots,k \\ q=1,\ldots,m}}$$

where

$$t_{uv} = \sum_{q=1}^{\ell} m_{uq} \cdot n_{qv}.$$

Example A3.6

Let
$$M = \begin{pmatrix} 2 & 3 & 9 \\ -1 & 4 & 0 \\ 5 & -3 & 6 \\ 4 & 4 & 1 \end{pmatrix}$$

and
$$N = \begin{pmatrix} -3 & 0 \\ 2 & 5 \\ -4 & -1 \end{pmatrix}.$$

Then $T = M \cdot N$ is well defined as a 4×2 matrix. We notice, for example, that

$$t_{11} = 2 \cdot (-3) + 3 \cdot 2 + 9 \cdot (-4) = -36$$

and

$$t_{32} = 5 \cdot 0 + (-3) \cdot 5 + 6 \cdot (-1) = -21.$$

Six other easy calculations of this kind yield that

$$M \cdot N = \begin{pmatrix} -36 & 6 \\ 11 & 20 \\ -45 & -21 \\ -8 & 19 \end{pmatrix}. \qquad \square$$

Definition A3.7

Let M be a $k \times k$ matrix. A matrix N is called the *inverse* of M if $M \cdot N = N \cdot M = I_k = I$, where

$$I = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ & & \cdots & \\ 0 & 0 & \cdots & 1 \end{pmatrix}.$$

When M has an inverse then it is called *invertible.* We denote the inverse by M^{-1}.

It follows immediately from the definition that, in order for a matrix to be a candidate for being invertible, it must be square.

Proposition A3.8

Let M be a $k \times k$ matrix with nonzero determinant. Then M is invertible and the elements of its inverse are given by

$$n_{ij} = \frac{(-1)^{i+j} \cdot \det M(i,j)}{\det M}.$$

Here $M(i,j)$ is the $(k-1) \times (k-1)$ matrix obtained by deleting the jth row and ith column from M.

Proof: This is a direct calculation. $\qquad \square$

Definition A3.9

If M is either a matrix or a vector, then the *transpose* ^tM of M is defined as follows: If the ijth entry M is m_{ij} then the ijth entry of ^tM is m_{ji}.

We will find the transpose notion useful primarily as notation. When we want to multiply a vector by a matrix, the multiplication will only make sense (in the language of matrix multiplication) after we have transposed the vector.

Proposition A3.10

If

$$\mathbf{u}^1 = (u_1^1, u_2^1, \ldots, u_k^1)$$

$$\mathbf{u}^2 = (u_1^2, u_2^2, \ldots, u_k^2)$$

$$\ldots$$

$$\mathbf{u}^k = (u_1^k, u_2^k, \ldots, u_k^k)$$

form a basis for $\mathbb{R}^k$ then let M be the matrix of the coefficients of these vectors and M^{-1} the inverse of M (which we know exists because the determinant of the matrix is nonzero). If $\boldsymbol{x} = (x_1, x_2, \ldots, x_k)$ is any element of $\mathbb{R}^k$ then

$$\boldsymbol{x} = a_1 \cdot \mathbf{u}^1 + a_2 \cdot \mathbf{u}^2 + \cdots + a_k \cdot \mathbf{u}^k,$$

where

$$(a_1, a_2, \ldots, a_k) = \boldsymbol{x} \cdot M^{-1}.$$

Proof: Let A be the vector of unknown coefficients $(a_1, a_2, \ldots, a_k)$. The system of equations that we need to solve to find $a_1, a_2, \ldots, a_k$ can be written in matrix notation as

$$A \cdot M = \boldsymbol{x}.$$

Applying the matrix M^{-1} to both sides of this equation (on the right) gives

$$(A \cdot M) \cdot M^{-1} = \boldsymbol{x} \cdot M^{-1}$$

or

$$A \cdot I = \boldsymbol{x} \cdot M^{-1}$$

or

$$A = \boldsymbol{x} \cdot M^{-1},$$

as desired. $\square$

The *standard basis* for $\mathbb{R}^k$ consists of the vectors

$$
\begin{aligned}
\mathbf{e}^1 &= (1, 0, \ldots, 0) \\
\mathbf{e}^2 &= (0, 1, \ldots, 0) \\
&\cdots \\
\mathbf{e}^k &= (0, 0, \ldots, 1)\,.
\end{aligned}
\tag{A3.11}
$$

If $\boldsymbol{x} = (x_1, x_2, \ldots, x_k)$ is any element of $\mathbb{R}^k$, then we may write

$$
\boldsymbol{x} = x_1\,\mathbf{e}^1 + x_2\,\mathbf{e}^2 + \cdots + x_k \cdot \mathbf{e}^k\,.
$$

In other words, the usual coordinates with which we locate points in k-dimensional space are the coordinates with respect to the special basis (A3.11). We write this basis as $\mathbf{e}^1, \mathbf{e}^2, \ldots, \mathbf{e}^k$.

If $\boldsymbol{x} = (x_1, x_2, \ldots, x_k)$ and $\mathbf{y} = (y_1, y_2, \ldots, y_k)$ are elements of $\mathbb{R}^k$ then we define

$$
\|\boldsymbol{x}\| = \sqrt{(x_1)^2 + (x_2)^2 + \cdots + (x_k)^2}
$$

and

$$
\boldsymbol{x} \cdot \mathbf{y} = x_1 y_1 + x_2 y_2 + \cdots + x_k y_k\,.
$$

Proposition A3.12   (The Schwarz Inequality)

If $\boldsymbol{x}$ and $\mathbf{y}$ are elements of $\mathbb{R}^k$ then

$$
|\boldsymbol{x} \cdot \mathbf{y}| \leq \|\boldsymbol{x}\|\|\mathbf{y}\|\,.
$$

Proof: Write out both sides and square. If all terms are moved to the right then the right side becomes a sum of perfect squares and the inequality is obvious. $\square$

Table of Notation

Notation	Section	Definition
$\mathbb{Q}$	1.1	the rational numbers
$\sup X$	1.1	supremum of X
$\operatorname{lub} X$	1.1	least upper bound of X
$\inf X$	1.1	infimum of X
$\operatorname{glb} X$	1.1	greatest lower bound of X
$\mathbb{R}$	1.1	the real numbers
$|x|$	1.1	absolute value
$|x + y| \leq |x| + |y|$	1.1	triangle inequality
$\mathcal{C}$	1.1AP	a cut
$\mathbb{C}$	1.2	the complex numbers
z	1.2	a complex number
i	1.2	the square root of -1
$\overline{z}$	1.2	complex conjugate
$|z|$	1.2	modulus of z
$e^{i\theta}$	1.2	complex exponential
$\{a_j\}$	2.1	a sequence
a_j	2.1	a sequence
a_{j_k}	2.2	a subsequence
$\liminf a_j$	2.3	limit infimum of a_j
$\limsup a_j$	2.3	limit supremum of a_j
a^j	2.4	a power sequence
e	2.4	Euler's number e
$\sum_{j=1}^{\infty} a_j$	3.1	a series
S_N	3.1	a partial sum
$\sum_{j=1}^{N} a_j$	3.1	a partial sum
$\sum_{j=1}^{\infty} (-1)^j b_j$	3.3	an alternating series
$j!$	3.4	j factorial
$\sum_{n=0}^{\infty} \sum_{j=0}^{n} a_j \cdot b_{n-j}$	3.5	the Cauchy product of series

Notation	Section	Definition
(a, b)	4.1	open interval
$[a, b]$	4.1	closed interval
$[a, b)$	4.1	half-open interval
$(a, b]$	4.1	half-open interval
U	4.1	an open set
F	4.1	a closed set
∂S	4.1	boundary of S
${}^{c}S$	4.1	complement of S
$\overline{S}$	4.2	closure of S
$\overset{\circ}{S}$	4.2	interior of S
$\{\mathcal{O}_\alpha\}$	4.3	an open cover
S_j	4.4	step in constructing the Cantor set
C	4.4	the Cantor set
$\lim_{E \ni x \to P} f(x)$	5.1	limit of f at P
ℓ	5.1	a limit
$f + g$	5.1	sum of functions
$f - g$	5.1	difference of functions
$f \cdot g$	5.1	product of functions
f/g	5.1	quotient of functions
$f \circ g$	5.2	composition of functions
f^{-1}	5.2	inverse function
$f^{-1}(W)$	5.2	inverse image of a set
$f(L)$	5.3	image of the set L
m	5.3	minimum for a function f
M	5.3	maximum for a function f
$\lim_{x \to P^-} f(x)$	5.4	left limit of f at P
$\lim_{x \to P^+} f(x)$	5.4	right limit of f at P
$f'(x)$	6.1	derivative of f at x
df/dx	6.1	derivative of f
$\text{Lip}_\alpha(I)$	6.3	space of Lipschitz-α functions
$C^{k,\alpha}(I)$	6.3	space of smooth functions of order k, α
$\mathcal{P}$	7.1	a partition
I_j	7.1	interval from the partition
Δ_j	7.1	length of I_j
$m(\mathcal{P})$	7.1	mesh of the partition
$\mathcal{R}(f, \mathcal{P})$	7.1	Riemann sum
$\int_a^b f(x)\, dx$	7.1	Riemann integral
$\int_b^a f(x)\, dx$	7.2	integral with reverse orientation
$\mathcal{U}(f, \mathcal{P}, \alpha)$	7.3	upper Riemann sum
$\mathcal{L}(f, \mathcal{P}, \alpha)$	7.3	lower Riemann sum
$I^*(f)$	7.3	upper integral of f

Notation	Section	Definition
$I_*(f)$	7.3	lower integral of f
$\int f \, d\alpha$	7.3	Riemann–Stieltjes integral
Vf	7.4	total variation of f
f_j	8.1	sequence of functions
$\{f_j\}$	8.1	sequence of functions
$\lim_{x \to s} f(x)$	8.2	limit of f as x approaches s
$\sum_{j=1}^{\infty} f_j(x)$	8.3	series of functions
$S_N(x)$	8.3	partial sum of a series o functions
$p(x)$	8.4	a polynomial
$\sum_{j=0}^{\infty} a_j(x-c)^j$	9.1	a power series
R_N	9.1	tail of the power series
ρ	9.2	radius of convergence
$f(x) =$ $\sum_{j=0}^{k} f^{(j)}(a)\frac{(x-a)^j}{j!} + R_{k,a}(x)$	9.2	Taylor expansion
$\exp(x)$	9.3	the exponential function
$\sin x$	9.3	the sine function
$\cos x$	9.3	the cosine function
$\text{Sin}\, x$	9.3	sine with restricted domain
$\text{Cos}\, x$	9.3	cosine with restricted domain
$\ln x$	9.4	the natural logarithm function
$dy/dx = F(x,y)$	10.1	first-order differential equation
$y(x) = y_0 + \int_{x_0}^{x} F(t, y(t)) \, dt$	10.1	integral equation equivalent of first order ODE
$y_{j+1}(x) = y_0 + \int_{x_0}^{x} F(t, y_j(t)) \, dt$	10.1	Picard iteration technique
$(j+1)a_{j+1} + (j-p)a_j = 0$	10.2	a recursion
$-2m(m-1) - m + 1 = 0$	10.2	indicial equation
c_n	11.1	Fourier coefficient
$\widehat{f}(n)$	11.2	nth Fourier coefficient
Sf	11.2	Fourier series
$S_N f$	11.2	partial sum of Fourier series
D_N	11.2	Dirichlet kernel
$\widehat{f}(\xi)$	11.3	Fourier transform of f
$C_0(\mathbb{R})$	11.3	continuous functions which vanish at ∞
a_n	11.4	Fourier cosine coefficient
b_n	11.4	Fourier sine coefficient
$\triangle$	11.4	Laplacian
$w(r,\theta) = \frac{1}{2}a_0$ $+ \sum_{j=1}^{\infty} r^j(a_j \cos j\theta + b_j \sin j\theta)$	11.4	solving the Dirichlet problem
$\mathbb{R}^k$	12.1	multidimensional Euclidean space

Notation	Section	Definition
$\boldsymbol{x} = (x_1, x_2, x_3)$	12.1	a point in multidimensional space
$B(\boldsymbol{x}, r)$	12.1	an (open) ball in multidimensional space
$\overline{B}(\boldsymbol{x}, r)$	12.1	a closed ball in multidimensional space
$\lim_{\boldsymbol{x} \to \mathbf{P}} f(\boldsymbol{x})$	12.1	limit in multidimensional space
$M_{\mathbf{P}}$	12.1	the derivative of f at $\mathbf{P}$
$\mathcal{R}_{\mathbf{P}}$	12.2	remainder term for the derivative
Jf	12.3	Jacobian matrix
$\mathbb{N}$	A1.1	the natural numbers
$\widehat{x}$	A1.1	successor
$Q(n)$	A1.1	inductive statement
$\binom{n}{k}$	A1.1	choose function
$\mathbb{Z}$	A1.2	the integers
$[(a, b)]$	A1.2	an integer
$\mathbb{Q}$	A1.3	the rational numbers
$[(c, d)]$	A1.3	a rational number
$\mathbf{A}$	A2.1	an atomic sentence
$\wedge$	A2.1	the connective "and"
$\vee$	A2.1	the connective "or"
$\sim$	A2.2	the connective "not"
$\Rightarrow$	A2.2	the connective "if-then"
$\{\ \}$	A2.5	a set
$\Leftrightarrow$	A2.3	the connective "if and only if"
iff	A2.3	the connective "if and only if"
$\forall$	A2.4	the quantifier "for all"
$\exists$	A2.4	the quantifier "there exists"
$\in$	A2.5	is an element of
$\subset$	A2.5	subset of
$\notin$	A2.5	is not an element of
$\cap$	A2.5	intersection
$\cup$	A2.5	union
$\emptyset$	A2.5	the empty set
$\setminus$	A2.5	set-theoretic difference
^{c}S	A2.5	complement of the set S

Notation	Section	Definition
(a, b)	A2.6	a relation
$f(x)$	A2.6	a function
$f \circ g$	A2.6	composition of functions
$\text{card}(A)$	A2.7	the cardinality of A
$|A|$	A2.7	the cardinality of A
E	A2.7	the even integers
O	A2.7	the odd integers
$\times$	A2.7	set-theoretic product
$\mathbf{u} = (u_1, \ldots, u_n)$	A3.1	a multidimensional vector
$\det$	A3.1	determinant
$M = (m_{pq})_{\substack{p=1,\ldots,k \\ q=1,\ldots,\ell}}$	A3.1	$k \times \ell$ matrix
$A \cdot B$	A3.1	matrix multiplication
$M(i, j)$	A3.1	deleted matrix
${}^t M$	A3.1	matrix transpose
$\boldsymbol{x} \cdot \mathbf{y}$	A3.1	vector dot product

GLOSSARY

Abel's convergence test   A test for convergence of series that is based on summation by parts.

absolutely convergent series   A series for which the absolute values of the terms form a convergent series.

absolute maximum   A number M is the absolute maximum for a function f if $f(x) \le f(M)$ for every x.

absolute minimum   A number m is the absolute minimum for a function f if $f(x) \ge f(m)$ for every x.

absolute value   Given a real number x, its absolute value is the distance of x to 0.

accumulation point   A point x is an accumulation point of a set S if every neighborhood of x contains infinitely many distinct elements of S.

algebraic number   A number which is the solution of a polynomial equation with integer coefficients.

alternating series   A series of real terms which alternate in sign.

alternating series test   If an alternating series has terms tending to zero then it converges.

"and"   The connective which is used for conjunction.

Archimedean Property   If a and b are positive real numbers then there is a positive integer n so that $na > b$.

atomic sentence   A sentence with a subject and a verb, and sometimes an object, but no connectives.

Bessel's inequality   An inequality for Fourier coefficients having the form $\sum_{n=-N}^{N} |\widehat{f}_n|^2 \leq \int_0^{2\pi} |f(t)|^2 \, dt$.

bijection   A one-to-one, onto function.

binomial expansion   The expansion, under multiplication, of the expression $(a + b)^n$.

Bolzano–Weierstrass Theorem   Every bounded sequence of real numbers has a convergent subsequence.

boundary of a set   The set of boundary points for the set.

boundary point   The point b is in the boundary of S if each neighborhood of b contains both points of S and points of the complement of S.

bounded above   A subset $S \subset \mathbb{R}$ is bounded above if there is a real number b such that $s \leq b$ for all $s \in S$.

bounded below   A subset $S \subset \mathbb{R}$ is bounded below if there is a real number c such that $s \geq c$ for all $s \in S$.

bounded sequence   A sequence a_j with the property that there is a number M so that $|a_j| \leq M$ for every j.

bounded set   A set S with the property that there is a number M with $|s| \leq M$ for every $s \in S$.

bounded variation   A function having bounded total oscillation.

Cantor set   A compact set which is uncountable, has zero length, is perfect, is totally disconnected, and has many other unusual properties.

cardinality   Two sets have the same cardinality when there is a one-to-one correspondence between them.

Cauchy Condensation Test A series of decreasing, nonnegative terms converges if and only if its dyadically condensed series converges.

Cauchy criterion   A sequence a_j is said to be Cauchy if, for each $\epsilon > 0$, there is an $N > 0$ so that, if $j, k > N$, then $|a_j - a_k| < \epsilon$.

Cauchy criterion for a series   A series satisfies the Cauchy criterion if and only if the sequence of partial sums satisfies the Cauchy criterion for a sequence.

Cauchy product   A means for taking the product of two series.

Cauchy's Mean Value Theorem   A generalization of the Mean Value Theorem that allows the comparison of two functions.

Chain Rule   A rule for differentiating the composition of functions.

change of variable   A method for transforming an integral by subjecting the domain of integration to a one-to-one function.

closed ball   The set of points at distance less than or equal to some $r > 0$ from a fixed point P.

closed set   The complement of an open set.

closure of a set   The set together with its boundary points.

common refinement of two partitions   The union of the two partitions.

compact set   A set E is compact if every sequence in E contains a subsequence that converges to an element of E.

comparison test for convergence   A series converges if it is majorized in absolute value by a convergent series.

comparison test for divergence   A series diverges if it majorizes a positive divergent series.

complement of a set   The set of points not in the set.

complete space   A space in which every Cauchy sequence has a limit.

complex conjugate   Given a complex number $z = x + iy$, the conjugate is the number $\overline{z} = x - iy$.

complex numbers   The set $\mathbb{C}$ of ordered pairs of real numbers equipped with certain operations of addition and multiplication.

composition   The composition of two functions is the succession of one function by the other.

conditionally convergent series   A series which converges, but not absolutely.

connected set   A set which cannot be separated by two disjoint open sets.

connectives   The words which are used to connect atomic sentences. These are "and," "or," "not," "if-then," and "if and only if."

continuity at a point   The function f is continuous at P if the limit of f at P equals the value of f at P. Equivalently, given $\epsilon > 0$, there is a $\delta > 0$ so that $|x - P| < \delta$ implies $|f(x) - f(P)| < \epsilon$.

continuous function   A function for which the inverse image of an open set is open.

continuously differentiable function   A function which has a derivative at every point, and so that the derivative function is continuous.

contrapositive   For a statement **"A implies B"**, the contrapositive statement is **"$\sim$ B implies $\sim$ A"**.

convergence of a sequence (of scalars)   A sequence a_j with the property that there is a limiting element ℓ so that, for any $\epsilon > 0$, there is a positive integer N so that, if $j > N$, then $|a_j - \ell| < \epsilon$.

convergence of a series   A series converges if and only if its sequence of partial sums converges.

converse   For a statement **"A implies B"**, the converse statement is **"B implies A"**.

cosine function   The function $\cos x = \sum_{j=0}^{\infty}(-1)^j x^{2j}/(2j)!$.

countable set   A set that has the same cardinality as the natural numbers.

Cramer's Rule   A device in linear algebra for solving systems of linear equations.

decreasing sequence   The sequence $\{a_j\}$ of real numbers is decreasing if $a_1 \geq a_2 \geq a_3 \geq \cdots$.

Dedekind cut   A rational halfline. Used to construct the real numbers.

de Morgan's Laws   The identities $^c(A \cup B) = {}^cA \cap {}^cB$ and $^c(A \cap B) = {}^cA \cup {}^cB$.

Density Property   If $c < d$ are real numbers then there is a rational number q with $c < q < d$.

denumerable set   A set that is either empty, finite, or countable.

derivative   The limit $\lim_{t \to x}(f(x) - f(t))/(t - x)$ for a function f on an open interval.

derived power series   The series obtained by differentiating a power series term by term.

determinant   The signed sum of products of elements of a matrix.

difference quotient   The quotient $(f(t) - f(x))/(t - x)$ for a function f on an open interval.

differentiable   A function that possesses the derivative at a point. This will be written differently in one variable and in several variables.

Dirichlet function   A function, taking only the values 0 and 1, which is highly discontinuous.

Dirichlet kernel   A kernel that represents the partial sum of a Fourier series. The kernel has the form $D_N(t) = (\sin(N + \frac{1}{2})t)/(\sin \frac{1}{2}t)$.

Dirichlet problem on the disc   The problem of finding a harmonic function on the disc with specified boundary values.

disconnected set   A set which can be separated by two disjoint open sets.

discontinuity of the first kind   A point at which a function f is discontinuous because the left and right limits at the point disagree.

discontinuity of the second kind   A point at which a function f is discontinuous because either the left limit or the right limit at the point does not exist.

diverge to infinity   A sequence with elements that become arbitrarily large.

domain of a function   See *function.*

domain of integration   The interval over which the integration is performed.

dummy variable   A variable whose role in an argument or expression is formal. A dummy variable can be replaced by any other variable with no logical consequences.

element of   A member of a given set.

empty set   The set with no elements.

equivalence classes   The pairwise disjoint sets into which an equivalence relation partitions a set.

equivalence relation   A relation that partitions the set in question into pairwise disjoint sets, called *equivalence classes.*

Euler's formula   The identity $e^{iy} = \cos y + i \sin y$.

Euler's number   This is the number $e = 2.71828\ldots$ which is known to be irrational, indeed transcendental.

exponential function   The function $\exp(z) = \sum_{j=0}^{\infty} z^j / j!$.

field   A system of numbers equipped with operations of addition and multiplication and satisfying eleven natural axioms.

finite-dimensional space   A linear space with a finite basis.

finite set   A set that can be put in one-to-one correspondence with a set of the form $\{1, 2, \ldots, n\}$ for some positive integer n.

finite subcovering   An open covering $\mathcal{U} = \{U_j\}_{j=1}^{k}$ is a finite subcovering of E if each element of $\mathcal{U}$ is an element of a larger covering $\mathcal{V}$.

first-order differential equation (ODE)   An equation of the form $dy/dx = F(x, y)$.

"for all"   The quantifier $\forall$ for making a statement about all objects of a certain kind.

Fourier coefficient   The coefficient $\widehat{f}(n) = (1/[2\pi]) \int_0^{2\pi} f(t)e^{-int}\, dt$ of the Fourier series for the function f.

Fourier series   A series of the form $f(t) \sim \sum_j c_j e^{ijt}$ which decomposes the function f as a sum of sines and cosines. We sometimes write $Sf \sim \sum_{j=-\infty}^{\infty} \widehat{f}(j)e^{ijt}$.

Fourier transform   Given a function f on the real line, its Fourier transform is $\widehat{f}(\xi) = \int_{\mathbb{R}} f(t)e^{it\xi}\, dt$.

function   A *function* from a set A to a set B is a relation f on A and B such that for each $a \in A$ there is one and only one pair $(a, b) \in f$. We call A the *domain* and B the *range* of the function.

Fundamental Theorem of Calculus   A result relating the values of a function to the integral of its derivative: $f(x) - f(a) = \int_a^x f'(t)\, dt$.

geometric series   This is a series of powers of a fixed base.

greatest lower bound   The real number c is the greatest lower bound for the set $S \subset \mathbb{R}$ if b is a lower bound and if there is no lower bound that is greater than c.

Green's function   A function $G(x, y)$ that is manufactured from the fundamental solution for the Laplacian and is useful in solving partial differential equations.

harmonic function   A function that is annihilated by the Laplacian.

Hausdorff space   A topological space in which distinct points p, q are separated by disjoint neighborhoods.

higher derivative   The derivative of a derivative.

Hilbert transform   The singular integral operator $f \mapsto \text{P.V.} \int f(t)/(x - t)\, dt$ which governs convergence of Fourier series and many other important phenomena in analysis.

i   The square root of -1 in the complex number system.

identity matrix   The square matrix with 1s on the diagonal and 0s in the other entries.

"if and only if"   The connective which is used for logical equivalence.

"if-then"   The connective which is used for implication.

image of a function   See *function*. The image of the function f is Image $f = \{b \in B : \exists a \in A \text{ such that } f(a) = b\}$.

image of a set   If f is a function then the image of E under f is the set $\{f(e) : e \in E\}$.

imaginary part   Given a complex number $z = x + iy$, its imaginary part is y.

implicit function theorem   A result that gives sufficient conditions, in terms of the derivative, on an equation of several variables to be able to solve for one variable in terms of the others.

increasing sequence   The sequence of real numbers a_j is increasing if $a_1 \leq a_2 \leq a_3 \leq \cdots$.

infimum   See *greatest lower bound*.

infinite set   A set is infinite if it is not finite.

initial condition   For a first-order differential equation, this is a side condition of the form $y(x_0) = y_0$.

integers   The natural numbers, the negatives of the natural numbers, and zero.

integral equation equivalent of a first-order ODE   An equation of the form $y(x) = y_0 + \int_{x_0}^{x} F(t, y(t)) \, dt$.

integration by parts   A device for integrating a product.

interior of a set   The collection of interior points of the set.

interior point   A point of the set S which has a neighborhood lying in S.

intermediate value theorem   The result that says that a continuous function does not skip values.

intersection of sets   The set of elements common to two or more given sets.

interval   A subset of the reals that contains all its intermediate points.

interval of convergence of a power series   An interval of the form $(c-\rho, c+\rho)$ on which the power series converges (uniformly on compact subsets of the interval).

inverse function theorem   A result that gives sufficient conditions, in terms of the derivative, for a function to be locally invertible.

inverse of a matrix   Given a square matrix A, we say that B is its inverse if $A \cdot B = B \cdot A = I$, where I is the identity matrix.

invertible matrix   A matrix that has an *inverse*.

irrational number   A real number which is not rational.

isolated point of a set   A point of the set with a neighborhood containing no other point of the set.

Jacobian matrix   The matrix of partial derivatives of a mapping from $\mathbb{R}^k$ to $\mathbb{R}^k$.

k times continuously differentiable   A function that has k derivatives, each of which is continuous.

Lambert W function   A transcendental function W with the property that any of the standard transcendental functions (sine, cosine, exponential, logarithm) can be expressed in terms of W.

Laplacian   The partial differential operator given by $\triangle = \partial^2/\partial x_1^2 + \partial^2/\partial x_2^2 + \cdots + \partial^2/\partial x_k^2$.

least upper bound   The real number b is the least upper bound for the set $S \subset \mathbb{R}$ if b is an upper bound and if there is no other upper bound that is less than b.

Least Upper Bound Property   The important defining property of the real numbers.

left limit   A limit of a function at a point P that is calculated with values of the function that are to the left of P.

Legendre's equation   The ODE $(1 - x^2)y'' - 2xy' + p(p + 1)y = 0$.

l'Hôpital's Rule   A rule for calculating the limit of the quotient of two functions in terms of the quotient of the derivatives.

limit   The value ℓ that a function approaches at a point of or an accumulation point P of the domain. Equivalently, given $\epsilon > 0$, there is a $\delta > 0$ so that $|f(x) - \ell| < \epsilon$ whenever $|x - P| < \delta$.

limit infimum   The least limit of any subsequence of a given sequence.

limit supremum   The greatest limit of any subsequence of a given sequence.

linear combination   If $\mathbf{v}_1$, $\mathbf{v}_2$, ..., $\mathbf{v}_k$ are vectors then a linear combination is an expression of the form $c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \cdots c_k\mathbf{v}_k$ for scalar coefficients c_j.

linearly dependent set   In a linear space, a set that is not linearly independent.

linearly independent set   In a linear space, a set that has no nontrivial linear combination giving 0.

linear operator   A function between linear spaces that satisfies the linearity condition $T(c\boldsymbol{x} + d\mathbf{y}) = cT(\boldsymbol{x}) + dT(\mathbf{y})$.

Lipschitz function   A function that satisfies a condition of the form $|f(s) - f(t)| \leq C|s - t|$ or $|f(s) - f(t)| \leq |s - t|^\alpha$ for $0 < \alpha \leq 1$.

local extrema   Either a local maximum or a local minimum.

local maximum   The point x is a local maximum for the function f if $f(x) \geq f(t)$ for all t in a neighborhood of x.

local minimum   The point x is a local minimum for the function f if $f(x) \leq f(t)$ for all t in a neighborhood of x.

logically equivalent   Two statements are logically equivalent if they have the same truth table.

logically independent   Two statements are logically independent if neither one implies the other.

lower bound   A real number c is a lower bound for a subset $S \subset \mathbb{R}$ if $s \geq c$ for all $s \in S$.

lower Riemann sum   A Riemann sum devised for defining the Riemann–Stieltjes integral.

Mean Value Theorem   If f is a continuous function on $[a, b]$, differentiable on the interior, then the slope of the segment connecting $(a, f(a))$ and $(b, f(b))$ equals the derivative of f at some interior point.

mesh of a partition   The maximum length of any interval in the partition.

metric   The distance function on a metric space.

metric space   A space X equipped with a distance function ρ.

$m \times n$ matrix   A matrix with m rows and n columns.

modulus   The modulus of a complex number $z = x + iy$ is $|z| = \sqrt{x^2 + y^2}$.

monotone sequence   A sequence that is either increasing or decreasing.

monotonically decreasing function   A function whose graph goes downhill when moving from left to right: $f(s) \geq f(t)$ when $s < t$.

monotonically increasing function   A function whose graph goes uphill when moving from left to right: $f(s) \leq f(t)$ when $s < t$.

monotonic function   A function that is either monotonically increasing or monotonically decreasing.

natural logarithm function   The inverse function to the exponential function.

natural numbers   The counting numbers $1, 2, 3, \ldots$.

necessary for   An alternative phrase for converse implication.

neighborhood of a point   An open set containing the point.

Neumann series   A series of the form $1/(1 - \alpha) = \sum_{j=0}^{\infty} \alpha^j$ for $|\alpha| < 1$.

Newton quotient   The quotient $(f(t) - f(x))/(t - x)$ for a function f on an open interval.

non-terminating decimal expansion   A decimal expansion for a real number that has infinitely many nonzero digits.

norm   The notion of distance on a normed linear space.

normed linear space   A linear space equipped with a norm that is compatible with the linear structure.

"not"   The connective which is used for negation.

one-to-one function   A function that takes different values at different points of the domain.

only if   An alternative phrase for implication.

onto   A function whose image equals its range.

open ball   The set of points at distance less than some $r > 0$ from a fixed point P.

open covering   A collection $\{U_\alpha\}_{\alpha \in A}$ of open sets is an open covering of a set S if $\cup_\alpha U_\alpha \supset S$.

Open Mapping Principle   The result that says that a bounded, surjective linear mapping is open.

open set   A set which contains a neighborhood of each of its points.

"or"   The connective which is used for disjunction.

ordered field   A field equipped with an order relation that is compatible with the field structure.

ordinary differential equation (ODE)   An equation involving a function of one variable and some of its derivatives.

partial derivative   For a function of several variables, this is the derivative calculated in just one variable, with the other variables held fixed.

partial sum of a Fourier series   The sum of the terms of a Fourier series having index between $-N$ and N.

partial sum of functions   The sum of the first N terms of a series of functions.

partial sum (of scalars)   The sum of the first N terms of a series of scalars.

partition of the interval $[a, b]$ A finite, ordered set of points $\mathcal{P} = \{x_0, x_1, x_2, \ldots, x_{k-1}, x_k\}$ such that

$$a = x_0 \leq x_1 \leq x_2 \leq \cdots \leq x_{k-1} \leq x_k = b\,.$$

Peano axioms   An axiom system for the natural numbers.

perfect set   A set which is closed and in which every point is an accumulation point.

Picard iteration technique   An iteration scheme for solving a first-order ODE using the steps $y_{j+1}(x) = y_0 + \int_{x_0}^{x} F(t, y_j(t))\, dt$.

Pinching Principle   A criterion for convergence of a sequence that involves bounding it below by a convergent sequence and bounding it above by another convergent sequence with the same limit.

pointwise convergence of a sequence of functions   A sequence f_j of functions converges pointwise if $f_j(x)$ convergence for each x in the common domain.

Poisson kernel   The reproducing kernel for harmonic functions.

polar form of a complex number   The polar form of a complex number z is $re^{i\theta}$, where r is the modulus of z and θ is the angle that the vector from 0 to z subtends with the positive x-axis.

power series expanded about the point c   A series of the form $\sum_{j=0}^{\infty} a_j (x - c)^j$.

power set   The collection of all subsets of a given set.

Principle of Induction   A proof technique for establishing a statement $Q(n)$ about the natural numbers.

quantifier   A logical device for making a quantitative statement. Our standard quantifiers are "for all" and "there exists."

radius of convergence of a power series   Half the length ρ of the interval of convergence.

range of a functon   See *function*.

rational numbers   Numbers which may be represented as quotients of integers.

Ratio Test for Convergence   A series converges if the limit of the sequence of quotients of summands is less than 1.

Ratio Test for Divergence   A series diverges if the limit of the sequence of quotients of summands is greater than 1.

real analytic function   A function with a convergent power series expansion about each point of its domain.

real numbers   An ordered field $\mathbb{R}$ containing the rationals $\mathbb{Q}$ so that every nonempty subset with an upper bound has a least upper bound.

real part   Given a complex number $z = x + iy$, its real part is x.

rearrangement of a series   A new series obtained by permuting the summands of the original series.

relation   A relation on sets A and B is a subset of $A \times B$.

remainder term for the Taylor expansion The term $R_{k,a}(x)$ in the Taylor expansion.

Riemann integrable A function for which the Riemann integral exists.

Riemann integral The limit of the Riemann sums.

Riemann–Lebesgue Lemma The result that says that the Fourier transform of an integrable function vanishes at infinity.

Riemann's lemma A result guaranteeing the existence of the Riemann–Stieltjes integral in terms of the proximity of the upper and lower Riemann sums.

Riemann–Stieltjes integral A generalization of the Riemann integral which allows measure of the length of the interval in the partition by a function α.

Riemann sum The approximate integral based on a partition.

right limit A limit of a function at a point P that is calculated with values of the function to the right of P.

Rolle's Theorem The special case of the Mean Value Theorem when $f(a) = f(b) = 0$.

Root Test for Convergence A series converges if the limit of the nth roots of the nth terms is less than one.

Root Test for Divergence A series is divergent if the limit of the nth roots of the nth terms is greater than one.

scalar An element of either $\mathbb{R}$ or $\mathbb{C}$.

Schroeder–Bernstein Theorem The result that says that if there is a one-to-one function from the set A to the set B and a one-to-one function from the set B to the set A then A and B have the same cardinality.

Schwarz inequality The inequality

$$|\mathbf{v} \cdot \mathbf{w}| \leq \|\mathbf{v}\| \|\mathbf{w}\|.$$

sequence of functions A function from $\mathbb{N}$ into the set of functions on some space.

sequence (of scalars)   A function from $\mathbb{N}$ into $\mathbb{R}$ or $\mathbb{C}$ or a metric space. We often denote the sequence by a_j.

series of functions   An infinite sum of functions.

series (of scalars)   An infinite sum of scalars.

set   A collection of objects.

setbuilder notation   The notation $\{x : P(x)\}$ for specifying a set.

set-theoretic difference   The set-theoretic difference $A \setminus B$ consists of those elements that lie in A but not in B.

set-theoretic isomorphism   A one-to-one, onto function.

set-theoretic product   If A and B are sets then their set-theoretic product is the set of ordered pairs (a, b) with $a \in A$ and $b \in B$.

sine function   The function $\sin x = \sum_{j=0}^{\infty} (-1)^j x^{2j+1}/(2j+1)!$.

smaller cardinality   The set A has smaller cardinality than the set B if there is a one-to-one mapping of A to B but none from B to A.

strictly monotonically decreasing function   A function whose graph goes strictly downhill when moving from left to right: $f(s) > f(t)$ when $s < t$.

strictly monotonically increasing function   A function whose graph goes strictly uphill when moving from left to right: $f(s) > f(t)$ when $s < t$.

subcovering   A covering which is a subcollection of a larger covering.

subfield   Given a field k, a subfield m is a subset of k which is also a field with the induced field structure.

subsequence   A sequence that is a subset of a given sequence with the elements occurring in the same order.

subset of   A subcollection of the members of a given set.

successor   The natural number which follows a given natural number.

suffices for   An alternative phrase for implication.

summation by parts   A discrete analogue of integration by parts.

supremum   See *least upper bound*.

Taylor's expansion   The expansion $f(x) = \sum_{j=0}^{k} f^{(j)}(a)\frac{(x-a)^j}{j!} + R_{k,a}(x)$ for a given function f.

terminating decimal   A decimal expansion for a real number that has only finitely many nonzero digits.

"there exists"   The quantifier $\exists$ for making a statement about some objects of a certain kind.

totally disconnected set   A set in which any two points can be separated by two disjoint open sets.

transcendental number   A real number which is not algebraic.

transpose of a matrix   Given a matrix $A = \{a_{ij}\}$, the transpose is the matrix obtained by replacing a_{ij} with a_{ji}.

triangle inequality   The inequality

$$|a + b| \le |a| + |b|$$

for real numbers.

truth table   An array which shows the possible truth values of a statement.

uncountable set   An infinite set that does not have the same cardinality as the natural numbers.

uniform convergence of a sequence of functions   The sequence f_j of functions converges uniformly to a function f if, given $\epsilon > 0$, there is an $N > 0$ so that, if $j > N$, then $|f_j(x) - f(x)| < \epsilon$ for all x.

uniform convergence of a series of functions   A series of functions such that the sequence of partial sums converges uniformly.

uniformly Cauchy sequence of functions   A sequence of functions f_j with the property that, for $\epsilon > 0$, there is an $N > 0$ so that, if $j, k > N$, then $|f_j(x) - f_k(x)| < \epsilon$ for all x in the common domain.

uniformly continuous   A function f is uniformly continuous if, for each $\epsilon > 0$, there is a $\delta > 0$ so that $|f(s) - f(t)| < \epsilon$ whenever $|s - t| < \delta$.

union of sets   The collection of objects that lie in any one of a given collection of sets.

universal set   The set of which all other sets are a subset.

upper bound   A real number b is an upper bound for a subset $S \subset \mathbb{R}$ if $s \leq b$ for all $s \in S$.

upper Riemann sum   A Riemann sum devised for defining the Riemann–Stieltjes integral.

Venn diagram   A pictorial device for showing relationships among sets.

Weierstrass Approximation Theorem   The result that any continuous function on $[0, 1]$ can be uniformly approximated by polynomials.

Weierstrass M-Test   A simple scalar test that guarantees the uniform convergence of a series of functions.

Weierstrass nowhere differentiable function   A function that is continuous on $[0, 1]$ that is not differentiable at any point of $[0, 1]$.

well defined   An operation on equivalence classes is well defined if the result is independent of the representatives chosen from the equivalence classes.

Zero Test   If a series converges then its summands tend to zero.

Bibliography

[**BOA1**] R. P. Boas, *A Primer of Real Functions*, Carus Mathematical Monograph No. 13, John Wiley & Sons, Inc., New York, 1960.

[**BIR**] G. Birkhoff and G.-C. Rota, *Ordinary Differential Equations*, John Wiley & Sons, New York, 1978.

[**BUC**] R. C. Buck, *Advanced Calculus*, 2d ed., McGraw-Hill Book Company, New York, 1965.

[**COL**] E. Coddington and N. Levinson, *Theory of Ordinary Differential Equations*, McGraw-Hill, New York, 1955.

[**DUS**] N. Dunford and J. Schwartz, *Linear Operators*, Interscience Publishers, New York, 1958–1971.

[**FED**] H. Federer, *Geometric Measure Theory*, Springer-Verlag, New York, 1969.

[**FOU**] J. Fourier, *The Analytical Theory of Heat*, G. E. Stechert & Co., New York, 1878.

[**HOF**] K. Hoffman, *Analysis in Euclidean Space*, Prentice Hall, Inc., Englewood Cliffs, NJ, 1962.

[**KAK**] S. Kakutani, Some characterizations of Euclidean space, *Japan. J. Math.* 16(1939), 93–97.

[**KAT**] Y. Katznelson, *Introduction to Harmonic Analysis*, John Wiley and Sons, New York, 1968.

[**KEL**] J. L. Kelley, Banach spaces with the extension property, *Trans. AMS* 72(1952), 323–326.

[**KOL**] A. N. Kolmogorov, *Grundbegriffe der Wahrscheinlichkeitsrechnung*, Springer-Verlag, Berlin, 1933.

[**KRA1**] S. G. Krantz, *The Elements of Advanced Mathematics*, 3rd ed., CRC Press, Boca Raton, FL, 2012.

[**KRA2**] S. G. Krantz, *A Panorama of Harmonic Analysis*, Mathematical Association of America, Washington, DC, 1999.

[**KRA3**] S. G. Krantz, *Partial Differential Equations and Complex Analysis*, CRC Press, Boca Raton, FL, 1992.

[**KRA4**] S. G. Krantz, *Handbook of Logic and Proof Techniques for Computer Scientists*, Birkhäuser, Boston, 2002.

[**KRA5**] S. G. Krantz, *Real Analysis and Foundations*, 3rd ed., CRC Press, Boca Raton, FL, 2013.

[**KRA6**] S. G. Krantz, *Function Theory of Several Complex Variables*, 2nd ed., American Mathematical Society, Providence, RI, 2001.

[**KRA7**] S. G. Krantz *Differential Equations: Theory, Technique, and Practice*, 2nd ed., Taylor & Francis/CRC Press, 2015.

[**KRA8**] S. G. Krantz, *Convex Analysis*, Taylor & Francis, Boca Raton, FL, 2015.

[**KRP**] S. G. Krantz and H. R. Parks, *A Primer of Real Analytic Functions*, 2nd ed., Birkhäuser Publishing, Boston, 2002.

[**LAN**] R. E. Langer, *Fourier Series: The Genesis and Evolution of a Theory*, Herbert Ellsworth Slaught Memorial Paper I, *Am. Math. Monthly* 54(1947).

[**LAX**] P. D. Lax, On the existence of Green's functions, *Proc. Amer. Math. Soc.* 3(1952), 526–531.

[**LOS**] L. Loomis and S. Sternberg, *Advanced Calculus*, Addison-Wesley, Reading, MA, 1968.

[**LUZ**] N. Luzin, The evolution of "Function", Part I, Abe Shenitzer, ed., *Am. Math. Monthly* 105(1998), 59–67.

[**NIV**] I. Niven, *Irrational Numbers*, Carus Mathematical Monograph No. 11, John Wiley & Sons, Inc., New York, 1956.

[**RES**] M. Reed and B. Simon, *Methods of Modern Mathematical Physics*, Academic Press, New York, 1972.

[**ROY**] H. Royden, *Real Analysis*, Macmillan, New York, 1963.

[**RUD1**] W. Rudin, *Principles of Mathematical Analysis*, 3rd ed., McGraw-Hill Book Company, New York, 1976.

[**RUD2**] W. Rudin, *Real and Complex Analysis*, McGraw-Hill Book Company, New York, 1966.

[**RUD3**] W. Rudin, *Functional Analysis*, McGraw-Hill, New York, 1973.

[**SOB**] A. Sobczyk, On the extension of linear transformations, *Trans. Amer. Math. Soc.* 55(1944), 153–169.

[**STG**] E. M. Stein and G. Weiss, *Introduction to Fourier Analysis on Euclidean Spaces*, Princeton University Press, Princeton, NJ, 1971.

[**STRO**] K. Stromberg, *An Introduction to Classical Real Analysis*, Wadsworth Publishing, Inc., Belmont, CA, 1981.

[**YOS**] K. Yosida, *Functional Analysis*, 6th ed., Springer-Verlag, New York, 1980.

Index